UNDERSTANDING AND ACCELERATING THE CONVERGENCE OF FEDERATED LORA

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated fine-tuning of large pretrained models often requires substantial communication and computation, making parameter-efficient methods such as LoRA attractive. A natural approach is FedAvg(LoRA), which combines FedAvg with LoRA. Despite its practical utility, the convergence theory of FedAvg(LoRA) remains limited: existing analyses typically assume bounded LoRA factors to ensure the smoothness of the factorized objective, an additional assumption not guaranteed by LoRA updates. More recent work on centralized LoRA shows that LoRA gradient descent achieves $O(\epsilon^{-4})$ complexity without this assumption. This raises the question of whether FedAvg(LoRA) can similarly achieve $O(\epsilon^{-4})$ communication complexity without this assumption. Unlike centralized LoRA, FedAvg(LoRA) performs multiple local steps that introduce client drift, and this drift can be amplified as the factor norms grow. We resolve this analytical challenge and establish an $O(\epsilon^{-6})$ communication upper bound under full client participation. We then prove a lower bound $\Omega(\epsilon^{-6})$ for multiple local steps, even when the stepsize $\eta_t \in (0, 1]$ is chosen adaptively from the optimization history. To accelerate this $\Omega(\epsilon^{-6})$ communication complexity, we introduce **FedStabLoRA**, which uses persistent client corrections to reduce client drift and stabilize training via a quadratic penalty. FedStabLoRA achieves $O(\epsilon^{-4})$ communication complexity without a bounded-factor assumption, and the $O(\epsilon^{-4})$ rate also holds under partial participation. Across seven language and vision benchmarks, FedStabLoRA achieves the best performance among the compared methods. Its advantage persists under severe data heterogeneity, more local training, and partial participation.

1 INTRODUCTION

Federated fine-tuning is often utilized to improve a shared pretrained model across different clients while keeping each client's training data local. FedAvg (McMahan et al., 2017) is a widely used federated learning method in which clients perform multiple local updates between server aggregations. For large pretrained models, however, updating and communicating the full model are costly. LoRA (Hu et al., 2022) reduces the number of trainable parameters by representing a weight update as $\Delta W = BA$, where only the low-rank factors A and B are optimized. Thus, in federated fine-tuning, LoRA is often combined with FedAvg as *FedAvg(LoRA)*: clients perform K local gradient steps on the LoRA factors, i.e., A and B, and the server averages the returned factors separately.

Analyzing FedAvg(LoRA) is challenging due to the LoRA reparameterization. Even when each client loss $F_i(W)$ is smooth in the weight variable W, the factored objective $J_i(B, A) = F_i(BA)$ is generally not globally smooth in (B, A) because the map $(B, A) \mapsto BA$ is bilinear (Sun et al., 2024). Concretely, if F_i is L-smooth in W, then for fixed A, the Lipschitz constant of $\nabla_B J_i$ with respect to B scales as $L\|A\|_2^2$, and analogously for A. Thus the local smoothness of J_i in factor space can grow with the current factor norms, so it is not uniformly bounded over (B, A). Existing convergence analyses of federated LoRA skip this difficulty by assuming that the LoRA factors remain uniformly bounded to ensure the smoothness of J_i (Chen et al., 2025; Zhang et al., 2026a). This is an additional assumption on the optimization trajectory and is not guaranteed by the LoRA updates. More recent centralized LoRA theory shows that this assumption is not necessary: original LoRA gradient descent can be analyzed under smoothness of the objective in W, and the latest analysis establishes an $O(\epsilon^{-4})$ complexity without this assumption (Wang et al., 2026). This raises

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.

our central question: *can FedAvg(LoRA) achieve the same $O(\epsilon^{-4})$ communication complexity without assuming bounded factors?*

Unlike centralized LoRA, FedAvg(LoRA) with multiple local steps introduces client drift (Karimireddy et al., 2020), and this drift can be amplified as the factor norms grow. With multiple local steps, heterogeneous clients follow different local factor trajectories, so subsequent gradients are evaluated at different local factor pairs. As the factor norms grow, the local factor gradients become more sensitive to these differences, which enlarges the error caused by client drift. We refer to this effect as *factor-amplified client drift*. Controlling this effect without assuming bounded factors is a central difficulty in analyzing FedAvg(LoRA).

We first show that FedAvg(LoRA) can be analyzed without bounded factors, but with multiple local steps ($K \geq 2$), its communication complexity degrades to $O(\epsilon^{-6})$, compared with the $O(\epsilon^{-4})$ complexity in the centralized setting. We use a fixed stepsize and show that it controls the local-update error induced by factor-amplified client drift without imposing a uniform bound on the factors.

We next prove that the $O(\epsilon^{-6})$ complexity remains tight under adaptive learning rates. At round t, the algorithm may choose any $\eta_t \in (0, 1]$ as a deterministic function of the optimization history available at the start of the round, with the same η_t applied to all local factor updates in that round. Specifically, for every fixed number of local steps $K \geq 2$ and sufficiently small ϵ, we construct a smooth two-client rank-one LoRA instance on which every such history-dependent learning-rate rule requires $\Omega(\epsilon^{-6})$ communication rounds. Thus, adapting the learning rate across rounds based on the optimization history cannot break the $\Omega(\epsilon^{-6})$ worst-case complexity barrier.

Finally, to accelerate this $\Omega(\epsilon^{-6})$ complexity, we introduce **FedStabLoRA**. Each client maintains a correction across rounds that reduces client drift while leaving the average update direction unchanged at the server iterate. The local update can additionally include a quadratic penalty that stabilizes the training procedure. FedStabLoRA achieves $O(\epsilon^{-4})$ communication complexity without assuming uniformly bounded factors. Moreover, for a fixed participation ratio q and a fixed number of clients, FedStabLoRA also achieves $O(\epsilon^{-4})$ communication complexity under uniform partial participation.

Contributions. Our main contributions are:

- **Upper bound for FedAvg(LoRA).** We prove an $O(\epsilon^{-6})$ communication complexity for full-participation FedAvg(LoRA) with multiple local steps ($K \geq 2$), without assuming uniformly bounded factors.
- **Matching lower bound.** For any fixed number of local steps $K \geq 2$, we prove that the adaptive learning rate requires $\Omega(\epsilon^{-6})$ communication rounds in the worst case, even when the learning rate is chosen adaptively based on the optimization history.
- **FedStabLoRA.** We introduce FedStabLoRA, which uses persistent client corrections to reduce factor-amplified client drift and an additional quadratic penalty that stabilizes the training. It achieves $O(\epsilon^{-4})$ communication complexity without a bounded-factor assumption. Under uniform partial participation, FedStabLoRA also achieves $O(\epsilon^{-4})$ complexity.
- **Empirical validation.** Across seven language and vision benchmarks, FedStabLoRA achieves the highest accuracy on every task. Its advantage persists under severe label skew, increased local training, and partial participation.

2 Related Work

Federated LoRA. Existing federated LoRA methods modify either the local LoRA updates or their server aggregation. FFA-LoRA fixes one LoRA factor and optimizes only the other (Sun et al., 2024), while FedEx-LoRA corrects the aggregation error caused by separately averaging the two factors (Singhal et al., 2025); we include both as baselines in our experiments. Other approaches modify the aggregation rule, align client factorizations, or adapt client-specific LoRA subspaces (Wang et al., 2024; Zhang et al., 2026b;a; Fang et al., 2026). In contrast, we study the standard multi-step FedAvg(LoRA) update, where each client optimizes both factors locally, and the server averages the returned factors separately.

Federated LoRA convergence theory. Recent analyses of federated LoRA either modify the update rule or impose assumptions on the LoRA factors. Chen et al. (2025) directly studies separate factor

averaging, but assumes uniformly bounded factors and bounded second moments of stochastic gradients. FSLoRA assumes bounded iterates and local smoothness in (A, B) (Fang et al., 2026), while Hou et al. (2025) studies independently sampled, sketched updates but assumes global smoothness in (A, B). We instead analyze FedAvg(LoRA) assuming only that the loss is smooth with respect to the weight variable W, without bounded factors or global smoothness in (A, B), and derive upper and lower communication-complexity bounds for the learning-rate class considered here.

LoRA convergence theory. Recent work has established convergence rates for centralized LoRA gradient descent. Mu & Klabjan (2025) analyze the original LoRA gradient-descent algorithm without assuming bounded factors or Lipschitz smoothness of the reparameterized objective, obtaining an $O(1/\log T)$ bound on the best-iterate squared factor-gradient norm. Wang et al. (2026) show that the same deterministic algorithm and stepsize reach factor-space ϵ-stationarity in $O(\epsilon^{-4})$ full-gradient evaluations. We use this $O(\epsilon^{-4})$ complexity as the comparison point for FedAvg(LoRA), whose complexity is measured in communication rounds.

Federated optimization. Glasgow et al. (2022) establish sharp upper and lower bounds for FedAvg/Local SGD on smooth stochastic objectives. SCAFFOLD uses control variates to reduce client drift (Karimireddy et al., 2020); FedProx adds a proximal penalty toward the server model (Li et al., 2020), while FedDyn uses dynamic regularization (Acar et al., 2021). FedStabLoRA uses a SCAFFOLD-style client correction and additionally adds a quadratic penalty that keeps the local LoRA factors close to the server factors. We prove that these corrected LoRA updates achieve $O(\epsilon^{-4})$ communication complexity without assuming bounded factors, including under partial participation. The connection to SCAFFOLD Option II is discussed in Appendix G.

3 Problem Setup

Federated objective. Consider N clients indexed by $[N] := \{1, \dots, N\}$. Let $\ell_i(W)$ denote client i's loss as a function of the full weight matrix, and let W_0 be the frozen pretrained weights. We define

$$F_i(X) := \ell_i(W_0 + X),$$

where $X \in \mathbb{R}^{m\times n}$ is the trainable weight update. Client i has aggregation weight $p_i > 0$, with $\sum_{i=1}^N p_i = 1$, and we write $p_{\min} := \min_{i\in[N]} p_i$. The global objective is

$$F(X) := \sum_{i=1}^N p_i F_i(X).$$

We use $\|\cdot\|_F$ for the Frobenius norm.

LoRA parameterization. LoRA parameterizes the update as

$$X = BA, \qquad B \in \mathbb{R}^{m\times r}, \qquad A \in \mathbb{R}^{r\times n},$$

and induces the factorized objectives

$$J_i(B, A) := F_i(BA), \qquad J(B, A) := \sum_{i=1}^N p_i J_i(B, A) = F(BA).$$

Their factor gradients are

$$\nabla_B J_i(B, A) = \nabla F_i(BA)A^\top, \qquad \nabla_A J_i(B, A) = B^\top \nabla F_i(BA).$$

Stacked factor variable. We stack the two factors as

$$V := \begin{bmatrix} B \\ A^\top \end{bmatrix} \in \mathbb{R}^{(m+n)\times r}, \qquad J_i(V) := J_i(B, A), \qquad J(V) := J(B, A).$$

We define

$$\nabla J_i(V) := \begin{bmatrix} \nabla_B J_i(B, A) \\ \nabla_A J_i(B, A)^\top \end{bmatrix}, \qquad \nabla J(V) := \sum_{i=1}^N p_i \nabla J_i(V).$$

Then $\|V\|_F^2 = \|B\|_F^2 + \|A\|_F^2$ and $\|\nabla J(V)\|_F^2 = \|\nabla_B J(B, A)\|_F^2 + \|\nabla_A J(B, A)\|_F^2$.

We call V ϵ-stationary if $\|\nabla J(V)\|_F \leq \epsilon$. For communication complexity, we use the best-iterate criterion: after T communication rounds, the algorithm reaches ϵ-stationarity if $\min_{0\leq t<T} \|\nabla J(V^t)\|_F \leq \epsilon$. Appendix H discusses its interpretation and its relation to stationarity in weight space.

Assumption 3.1 (Weight-space smoothness)**.** There exists $L > 0$ such that, for every $i \in [N]$ and all $X, Y \in \mathbb{R}^{m\times n}$,

$$\|\nabla F_i(X) - \nabla F_i(Y)\|_F \leq L\|X - Y\|_F.$$

Assumption 3.2 (Lower-bounded global objective)**.** There exists $F_\star > -\infty$ such that $F(X) \geq F_\star$ for all $X \in \mathbb{R}^{m\times n}$.

Assumption 3.3 (Bounded gradient heterogeneity)**.** There exists $\zeta \geq 0$ such that, for all $X \in \mathbb{R}^{m\times n}$,

$$\sum_{i=1}^{N} p_i \|\nabla F_i(X) - \nabla F(X)\|_F^2 \leq \zeta^2.$$

For the initialization V^0, let $\Delta_0 := J(V^0) - F_\star$.

FedAvg(LoRA). We first study full participation with exact local gradients. At round t, each client initializes $V_{i,0}^t = V^t$ and performs $K \geq 1$ local steps:

$$V_{i,e+1}^t = V_{i,e}^t - \eta_t \nabla J_i(V_{i,e}^t), \qquad e = 0, \ldots, K-1,$$
$$V^{t+1} = \sum_{i=1}^{N} p_i V_{i,K}^t.$$

A single stepsize η_t is used across all clients, local steps, and both LoRA factors in round t. The server averages the returned B and A factors separately, rather than the products B_iA_i. When $K = 1$, the update becomes $V^{t+1} = V^t - \eta_t \nabla J(V^t)$, i.e., centralized gradient descent on J.

4 Convergence of FedAvg(LoRA)

We study the convergence of FedAvg(LoRA) with K local steps per communication round, without assuming bounded LoRA factors. Under client heterogeneity and multiple local steps, we obtain an $O(\epsilon^{-6})$ communication complexity. This bound reduces to the centralized $O(\epsilon^{-4})$ rate when $K = 1$ or the clients are homogeneous. We further show that, for every fixed $K \geq 2$, the $O(\epsilon^{-6})$ complexity is unavoidable in the worst case.

4.1 Upper Bound

We show that full-participation FedAvg(LoRA) converges without assuming bounded LoRA factors.

Theorem 4.1 (FedAvg(LoRA) upper bound)**.** *Under Assumptions 3.1–3.3, there exists $a_0 \in (0, 1]$, depending only on $L, p_{\min}, \Delta_0, \|V^0\|_F^2$, and a fixed upper bound on ζ, such that the following holds. For any number of local steps $K \geq 1$, any number of communication rounds $T \geq 1$, and any fixed $a \in (0, a_0]$, full-participation FedAvg(LoRA) with local stepsize*

$$\eta = \frac{a}{K\left[\sqrt{T} + \big((1-1/K)\zeta\big)^{1/3} T^{2/3}\right]} \tag{1}$$

satisfies

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla J(V^t)\|_F^2 \leq \frac{4\Delta_0 + 1}{a}\left(\frac{1}{\sqrt{T}} + \frac{\big((1-1/K)\zeta\big)^{1/3}}{T^{1/3}}\right). \tag{2}$$

Consequently, FedAvg(LoRA) reaches ϵ-stationarity within $O\big(\epsilon^{-4} + (1-1/K)\zeta\,\epsilon^{-6}\big)$ communication rounds.

The stepsize is chosen before training and remains fixed across communication rounds, while the factor norms are controlled by the analysis rather than assumed bounded. When $K = 1$ or there is no client heterogeneity ($\zeta = 0$), the second term vanishes and FedAvg(LoRA) achieves $O(\epsilon^{-4})$ communication complexity. When $K > 1$ and $\zeta > 0$, multiple local steps together with client heterogeneity introduce the additional $O((1-1/K)\zeta\epsilon^{-6})$ term. The proof is given in Appendix B.

4.2 LOWER BOUND

The $O(\epsilon^{-6})$ upper bound raises the question of whether an adaptive stepsize schedule, possibly depending on the entire optimization history, can reduce the number of communication rounds needed to reach an ϵ-stationary point. For $K \geq 2$, the following lower bound rules this out in the worst case.

Theorem 4.2 (Matching lower bound)**.** *For every fixed number of local steps $K \geq 2$, there exist constants $c_K > 0$ and $\epsilon_0(K) > 0$ such that, for every $0 < \epsilon \leq \epsilon_0(K)$, there exists a two-client, rank-one LoRA instance with $p_1 = p_2 = 1/2$ satisfying Assumptions 3.1–3.3 for which every stepsize sequence $(\eta_t)_{t\geq 0}$ with $0 < \eta_t \leq 1$ requires at least $c_K \epsilon^{-6}$ communication rounds to reach an ϵ-stationary iterate. Moreover, L, ζ, Δ_0, and $\|V^0\|_F$ are bounded by constants independent of K and ϵ.*

The hard instance is independent of the stepsize sequence, so the lower bound also applies when η_t is chosen deterministically from the preceding optimization history. Thus, for every fixed $K \geq 2$, the lower bound matches the $O(\epsilon^{-6})$ upper bound in Theorem 4.1, up to constants depending on K.

Proof idea. The construction uses two clients with aggregation weights $p_+ = p_- = 1/2$:

$$F_+(X) = F_0(X) + Q(X), \qquad F_-(X) = F_0(X) - Q(X).$$

Hence $F(X) = \frac{1}{2}F_+(X) + \frac{1}{2}F_-(X) = F_0(X)$, while the client gradients are $\nabla F_0(X) \pm \nabla Q(X)$. Thus Q disappears from the global objective but still changes the local updates. Starting from the same server factors, the two clients produce different factor pairs after the first local step. Their subsequent gradients are therefore evaluated at different factor pairs before the server averages the returned factors. This produces factor-amplified client drift.

To track the server's progress, let $x_t := \langle e_1 e_1^\top, B^t A^t \rangle = (B^t A^t)_{11}$. We choose F_0 and Q so that the server must cross $\Theta_K(\epsilon^{-2})$ separated intervals of x_t before reaching an ϵ-stationary point. Inside each interval, one communication round can increase x_t by at most $O_K(\epsilon^4)$, and the construction prevents a round from jumping over the interval. Since each interval has width bounded below independently of ϵ, crossing one interval requires $\Omega_K(\epsilon^{-4})$ rounds. Crossing all $\Theta_K(\epsilon^{-2})$ intervals therefore requires $\Omega_K(\epsilon^{-6})$ rounds. The complete construction and proof are given in Appendix C.

5 FEDSTABLORA

Section 4 shows that, for $K \geq 2$, changing the stepsize schedule alone cannot improve the $\Theta(\epsilon^{-6})$ worst-case communication complexity of FedAvg(LoRA). FedStabLoRA instead modifies the local update using a persistent client-specific correction inspired by the control-variate mechanism in SCAFFOLD (Karimireddy et al., 2020) and an additional quadratic penalty loss.

5.1 METHOD

Client i keeps a correction state Y_i^t, and the server maintains its weighted average $\bar{Y}^t := \sum_{i=1}^N p_i Y_i^t$. Both are initialized to zero. We define the centered correction $C_i^t := Y_i^t - \bar{Y}^t$. Since $\sum_i p_i C_i^t = 0$, $\sum_{i=1}^N p_i \big[\nabla J_i(V^t) - C_i^t\big] = \nabla J(V^t)$. At the start of each round, the weighted average of the corrected client gradients equals the global gradient at V^t.

At round t, each participating client starts from $V_{i,0}^t = V^t$ and uses the same correction C_i^t throughout its K local steps:

$$V_{i,e+1}^t = V_{i,e}^t - \eta\Big[\nabla J_i(V_{i,e}^t) - C_i^t + \mu(V_{i,e}^t - V^t)\Big], \qquad e = 0, \ldots, K-1. \tag{3}$$

The term $\mu(V_{i,e}^t - V^t)$ is the gradient of the quadratic penalty $\frac{\mu}{2}\|V_{i,e}^t - V^t\|_F^2$, which pulls the local iterate toward the server iterate V^t.

After the local steps, client i returns $D_i^t := V_{i,K}^t - V^t$ and sets $Y_i^{t+1} := Y_i^t - D_i^t/s$. The choice of s is motivated by

$$D_i^t \approx -s\big[\nabla J_i(V^t) - C_i^t\big].$$

Under this relation, $-D_i^t/s$ approximates the corrected client gradient at the start of round and can therefore be used to update the correction state. We use the same (η, μ, s) across clients and rounds.

Algorithm 1 FedStabLoRA

Require: Initial factors V^0; rounds T; local steps $K \geq 1$; parameters $\eta > 0$, $\mu \geq 0$, and $s > 0$

Initialize $Y_i^0 \leftarrow 0$ for all i and $\bar{Y}^0 \leftarrow 0$
for $t = 0, \ldots, T-1$ **do**
  Choose participating clients $S_t \subseteq [N]$
  Send $(V^t, \bar{Y}^t)$ to each $i \in S_t$
  for all $i \in S_t$ **in parallel do**
    $C_i^t \leftarrow Y_i^t - \bar{Y}^t, \quad V_{i,0}^t \leftarrow V^t$
    for $e = 0, \ldots, K-1$ **do**
      $V_{i,e+1}^t \leftarrow V_{i,e}^t - \eta\big[\nabla J_i(V_{i,e}^t) - C_i^t + \mu(V_{i,e}^t - V^t)\big]$
    end for
    $D_i^t \leftarrow V_{i,K}^t - V^t, \quad Y_i^{t+1} \leftarrow Y_i^t - D_i^t/s$
    Send D_i^t to the server
  end for
  $Y_i^{t+1} \leftarrow Y_i^t$ for $i \notin S_t$
  $V^{t+1} \leftarrow V^t + \dfrac{\sum_{i \in S_t} p_i D_i^t}{\sum_{i \in S_t} p_i}$
  $\bar{Y}^{t+1} \leftarrow \bar{Y}^t - \dfrac{1}{s} \sum_{i \in S_t} p_i D_i^t$
end for

For the participating clients S_t, the server updates $V^{t+1} := V^t + \frac{\sum_{i \in S_t} p_i D_i^t}{\sum_{i \in S_t} p_i}$ and $\bar{Y}^{t+1} := \bar{Y}^t - \frac{1}{s} \sum_{i \in S_t} p_i D_i^t$. Clients outside S_t keep $Y_i^{t+1} = Y_i^t$, so $\bar{Y}^{t+1} = \sum_i p_i Y_i^{t+1}$ remains exact after every round. Under full participation, $S_t = [N]$. Under partial participation, we take $p_i = 1/N$ and sample M clients uniformly without replacement. Algorithm 1 summarizes the overall procedure.

5.2 Convergence under Full Participation

We first consider full participation, where all clients participate in every round. We analyze the parameterization used by FedStabLoRA in our experiments.

Parameter selection. For $\mu \geq 0$ and $0 \leq \eta\mu \leq 1$, define

$$s := \eta \sum_{k=0}^{K-1} (1 - \eta\mu)^k. \tag{4}$$

If the local gradient is frozen at V^t, the returned update is exactly $D_i^t = -s[\nabla J_i(V^t) - C_i^t]$. When $\mu = 0$, this reduces to $s = K\eta$.

Theorem 5.1 (FedStabLoRA under full participation)**.** *Under Assumptions 3.1–3.3, there exist constants $c_1, C_2 > 0$, depending only on $L, \zeta, p_{\min}, \Delta_0$, and $\|V^0\|_F^2$, such that the following holds. For any number of local steps $K \geq 1$ and any communication horizon $T \geq 1$, using (4) with $s = c_1/\sqrt{T}$ gives*

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla J(V^t)\|_F^2 \leq \frac{C_2}{\sqrt{T}}. \tag{5}$$

Consequently, FedStabLoRA reaches an ϵ-stationary iterate within $O(\epsilon^{-4})$ communication rounds.

FedStabLoRA therefore achieves the $O(\epsilon^{-4})$ communication complexity of the $K = 1$ case even when $K \geq 2$. The guarantee does not deteriorate with the number of local steps K. In particular, setting $\mu = 0$ and $\eta = c_1/(K\sqrt{T})$ satisfies the same guarantee. Thus a positive quadratic penalty is not required for this theoretical rate. Empirically, however, we find that the penalty stabilizes the corrected LoRA updates and allows the full method to use a larger learning rate; we study this effect in Section 6.3.

Proof idea. Section 5.1 motivates the relation $-D_i^t/s \approx \nabla J_i(V^t) - C_i^t$. After updating Y_i^t and subtracting the weighted average of the updated client states, the new correction satisfies

$C_i^{t+1} \approx \nabla J_i(V^t) - \nabla J(V^t)$. Thus, the correction used in the next round approximately removes the difference between each client gradient and the global gradient. We then construct a Lyapunov function combining the objective gap and the correction error, and show that it decreases in each round. Summing the resulting descent inequality over all rounds gives $\sum_{t<T} s\|\nabla J(V^t)\|_F^2 = O(1)$. With $s = \Theta(T^{-1/2})$, the claimed $O(T^{-1/2})$ rate follows. The complete proof is given in Appendix D.

5.3 Convergence under Partial Participation

We next consider uniform partial participation. In each round, M of the N clients are sampled uniformly without replacement. Only sampled clients update their correction states.

Theorem 5.2 (FedStabLoRA under partial participation)**.** *Under Assumptions 3.1–3.3, there exist constants $c_3, C_4 > 0$, depending only on L, ζ, Δ_0, and $\|V^0\|_F^2$, such that the following holds. For any $N \geq 1$, $M \in \{1, \ldots, N\}$, $K \geq 1$, and $T \geq 1$, using (4) with $s = c_3 M/(N^{3/2}\sqrt{T})$ gives*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla J(V^t)\|_F^2 \leq \frac{C_4 N^{3/2}}{M\sqrt{T}}, \tag{6}$$

where the expectation is over client sampling. Consequently, $\min_{0\leq t<T} \mathbb{E}\|\nabla J(V^t)\|_F^2 \leq \epsilon^2$ after $O\big(N^3 M^{-2}\epsilon^{-4}\big)$ communication rounds.

For a fixed number of clients N and participation ratio $q = M/N$, FedStabLoRA achieves $O(\epsilon^{-4})$ complexity, with no deterioration as the number of local steps K increases. The same guarantee holds for $\mu = 0$ with $\eta = c_3 M/(KN^{3/2}\sqrt{T})$. The complete proof is given in Appendix E.

6 Experiments

We evaluate FedStabLoRA across language and vision tasks, and then study how it behaves as client heterogeneity, local training, client participation, and model size change.

Datasets and federated setting. Our main evaluation covers five language benchmarks: MNLI-m, QQP, QNLI, and SST-2 from GLUE (Wang et al., 2019), and AG News (Zhang et al., 2015). We additionally evaluate on CIFAR-100 (Krizhevsky, 2009) and DomainNet (Peng et al., 2019). Unless otherwise stated, we use 10 clients with full participation and Dirichlet label skew with concentration parameter $\alpha = 0.1$. Every client receives the same number of training examples. Each run lasts 50 communication rounds with a local-training budget equivalent to $E = 2$ passes over each client's data. We report this budget as $E = 2$ local epochs. The implementation uses the corresponding fixed number $K = \lceil En/B \rceil$ of local SGD updates, where n is the client data size and B is the batch size; K is also the local-step count used in the theory. Appendix F.3 reports the K for every setting.

Models and training. For the language tasks, we fine-tune RoBERTa-base (Liu et al., 2019); for the vision tasks, we use a ViT-small model (Dosovitskiy et al., 2021) pretrained on ImageNet-21k. All experiments use constant-step SGD. For LoRA (Hu et al., 2022), we use rank 4, scaling parameter 8. FedStabLoRA uses the correction scale in (4). The correction scale s is computed from the learning rate, μ, and the number K of local SGD steps; it is not tuned separately. Implementation details are provided in Appendix F.

Baselines and evaluation. We compare with FedAvg (McMahan et al., 2017), FedProx (Li et al., 2020), FedDyn (Acar et al., 2021), and SCAFFOLD (Karimireddy et al., 2020), as well as the federated LoRA methods FedEx-LoRA (Singhal et al., 2025) and FFA-LoRA (Sun et al., 2024). Within each benchmark, all methods use the same backbone, optimizer family, batch size, local-update budget, and communication round budget. We tune the local learning rate and any method-specific coefficient separately on the tuning split. We report accuracy for all seven benchmarks as mean $\pm$ standard deviation over three seeds. Appendix F gives the complete configurations.

6.1 Overall Performance

Table 1 shows a consistent cross-task advantage under Dirichlet label skew with $\alpha = 0.1$, across both language and vision tasks and against both general federated-learning and federated-LoRA baselines.

Table 1: Main benchmark comparison under Dirichlet label skew with $\alpha = 0.1$ and full client participation. FedStabLoRA achieves the highest mean accuracy on all seven language and vision benchmarks. Results are mean $\pm$ standard deviation over three seeds; best means are shown in bold.

Method	MNLI-m	QQP	QNLI	SST-2	AGNews	CIFAR100	DomainNet
FedAvg	76.27±2.36	82.15±1.32	88.26±1.37	90.57±1.04	85.11±1.69	86.62±0.51	76.25±1.88
FedProx	81.12±0.76	86.89±0.79	88.95±0.96	91.02±0.32	89.85±0.76	88.59±0.49	77.24±0.17
FedDyn	82.71±1.02	86.48±0.92	89.07±0.79	90.88±1.16	90.47±0.16	88.48±0.37	76.78±0.28
SCAFFOLD	81.25±1.64	87.55±1.06	90.47±1.87	91.83±0.35	90.73±0.36	88.39±0.20	76.57±0.60
FedEx-LoRA	77.40±1.40	83.83±1.64	88.62±1.09	90.15±1.52	88.17±1.17	87.64±0.54	75.26±0.19
FFA-LoRA	76.28±1.47	81.13±1.88	88.04±1.14	89.95±1.42	87.68±1.64	87.62±0.26	76.67±0.01
FedStabLoRA	**85.07**±0.65	**90.15**±0.71	**91.59**±0.90	**93.48**±0.50	**92.12**±0.61	**89.86**±0.23	**78.42**±0.16

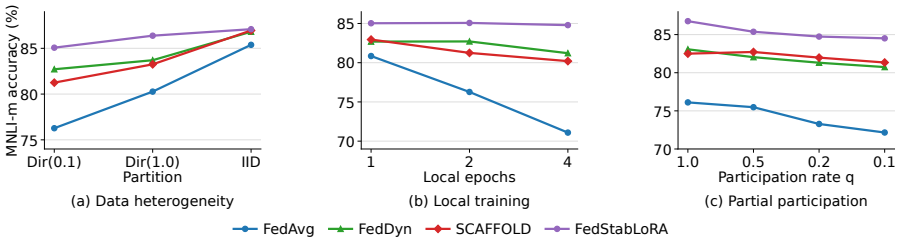

Figure 1: Effects of data heterogeneity, local training, and partial participation on MNLI-m. (a) Varying client-data heterogeneity from Dirichlet label skew ($\alpha = 0.1$ and 1.0) to IID, with $E = 2$. (b) Varying the local-training budget $E \in \{1, 2, 4\}$ under Dirichlet label skew with $\alpha = 0.1$. (c) Varying the participation rate q under Dirichlet label skew with $\alpha = 0.1$. Panels (a)–(b) use 10 clients, full participation, and 50 communication rounds; panel (c) uses 50 clients, $E = 1$, and 200 communication rounds. Results are averaged over three seeds.

FedStabLoRA achieves the highest mean accuracy on all seven benchmarks, improving over the strongest competing method on each task by 1.12–2.60 percentage points.

Scaling to larger backbones. We further evaluate Qwen2.5-1.5B, 3B, and 7B on MNLI-m under $\alpha = 0.1$. FedStabLoRA achieves the highest mean performance at all three model sizes, while its margin over the strongest baseline narrows at 7B. Prior work reports that larger pretrained transformers can be more robust to client data heterogeneity (Zhang et al., 2023). This may partly explain the smaller margin at 7B. These results show that the benefit of FedStabLoRA persists as the backbone scales, even as the pretrained model becomes more capable.

6.2 Heterogeneity, Local Training, and Partial Participation

Data heterogeneity and local training. Figure 1(a) shows that the performance gap between FedStabLoRA and the baselines widens as client heterogeneity increases. Under severe label skew ($\alpha = 0.1$), FedStabLoRA substantially outperforms FedAvg and SCAFFOLD; as the client distributions become more homogeneous, the gap substantially narrows under IID data. This trend supports our motivation that FedStabLoRA is particularly useful when client heterogeneity induces stronger factor-amplified client drift.

Figure 1(b) shows that increasing the amount of local training leads to a clearer advantage for FedStabLoRA. As E increases, the performance of FedAvg, FedDyn, and SCAFFOLD degrades, while FedStabLoRA remains stable across the tested local-training budgets. This observation is consistent with our analysis: under heterogeneous client objectives, additional local updates increase client drift and make drift correction more important. Exact results are reported in Appendix F.7.

Partial participation. Figure 1(c) shows that FedStabLoRA remains effective as the client participation rate decreases. Across all tested participation rates, it achieves the highest mean accuracy among the compared methods. FedStabLoRA is also more communication-efficient: at a common target of 78% accuracy, it requires approximately 28–52% fewer communication rounds than the fastest competing method. These results show that the advantage of FedStabLoRA persists under

client subsampling, both in final performance and in the number of communication rounds required to reach a given accuracy. Complete results are reported in Appendix F.8.

6.3 Ablation and Stability Analysis

We next examine the roles of the correction term and quadratic penalty in FedStabLoRA. Table 2 reports a component ablation on MNLI-m under Dirichlet label skew with $\alpha = 0.1$.

Table 2: Component ablation under LoRA and full fine-tuning on MNLI-m with Dirichlet label skew ($\alpha = 0.1$). The $s = K\eta$ variant uses the same correction and quadratic penalty as FedStabLoRA but replaces its correction scale with $s = K\eta$. Results are mean $\pm$ standard deviation over three seeds.

Parameterization	Base	+Penalty	+Correction†	+Both ($s = K\eta$)	FedStabLoRA
LoRA ($r = 4$)	75.62±1.63	80.05±0.34	81.25±1.64	82.99±0.47	**85.07±0.65**
Full fine-tuning	82.43±0.82	82.83±0.67	86.16±0.53	86.08±0.65	**86.52±0.93**

†The LoRA correction-only variant diverges at the FedStabLoRA learning rate $\eta = 0.3$. Its learning rate is therefore retuned on the tuning split. All other ablation variants use the selected FedStabLoRA configuration.

Component ablation. Under LoRA, both the correction and the quadratic penalty improve over the base method when used individually. Combining the two gives a further improvement, and replacing $s = K\eta$ with the proposed correction scale yields the best mean accuracy. This shows that the improvement of FedStabLoRA cannot be attributed to a single component.

Optimization stability. The quadratic penalty is particularly important for stabilizing the corrected LoRA update. Without the penalty, the correction-only variant diverges at the learning rate used by FedStabLoRA ($\eta = 0.3$) and must be retuned with a smaller learning rate. With the penalty, the corrected update remains stable at $\eta = 0.3$. Thus, although $\mu > 0$ is not required for the $O(\epsilon^{-4})$ guarantee in Section 5.2, empirically it allows the corrected LoRA update to use a larger learning rate, leading to faster finite-round optimization. Details of the ablation and retuning protocol are provided in Appendix F.4, and learning-rate sensitivity results are reported in Table 15.

LoRA versus full fine-tuning. The same behavior is not observed under full fine-tuning. There, the correction alone already accounts for most of the improvement over the base method, and adding the penalty changes the final accuracy only slightly. This contrast suggests that the stabilizing effect of the quadratic penalty is especially useful when the correction is applied to the LoRA factors.

6.4 Communication, Memory, and Runtime Overhead

FedStabLoRA maintains one additional P-dimensional state for each client and one on the server, where P is the number of trainable parameters. In each round, a participating client downloads the current model and the aggregated correction state ($2P$ in total) and uploads its local update ($1P$). This gives the same memory and communication cost as SCAFFOLD, and $1.5\times$ the per-round communication of FedAvg-LoRA. Using the same H200 hardware and training configuration for all methods, FedStabLoRA takes $1.013\times$ the per-round time of FedAvg-LoRA and increases peak memory by 36 MB (0.42%) in the 10-client simulator. The main additional overhead therefore comes from storing and communicating the correction state, while the local computation cost remains close to that of FedAvg-LoRA. Appendix F.11 provides the complete overhead and measurement setup.

7 Conclusion

We studied FedAvg(LoRA) without assuming bounded LoRA factors or global smoothness in the factor variables (A, B). For FedAvg(LoRA), we proved an $O(\epsilon^{-6})$ communication upper bound and an $\Omega(\epsilon^{-6})$ lower bound for $\eta_t \in (0, 1]$ when clients perform multiple local updates per round. We then introduced FedStabLoRA, which uses persistent client corrections and an additional quadratic penalty to reduce factor-amplified client drift and achieves $O(\epsilon^{-4})$ communication complexity, including under partial participation. Across language and vision tasks, FedStabLoRA achieves higher accuracy than the compared methods under strong label skew, more local training, and partial participation.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing, identifying potentially relevant literature, and code generation. All citations included in the paper were independently located and verified by the authors using DBLP rather than taken directly from AI-generated outputs. AI-generated code was reviewed, tested, and validated by the authors before being used in the experiments. The authors take full responsibility for the manuscript, citations, code, and reported results.

REPRODUCIBILITY STATEMENT

Complete proofs of all theoretical results are provided in the appendix. Experimental settings, data splits, model and optimization configurations, hyperparameter selection procedures, and additional results are documented in Appendix F.

REFERENCES

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N. Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=B7v4QMR6Z9w.

Xin Chen, Shuaijun Chen, Omid Tavallaie, Nguyen Tran, Shuhuang Xiang, and Albert Y. Zomaya. Convergence analysis of aggregation-broadcast in lora-enabled distributed fine-tuning. *CoRR*, abs/2508.01348, 2025. doi: 10.48550/ARXIV.2508.01348. URL https://doi.org/10.48550/arXiv.2508.01348.

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 *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Wenzhi Fang, Dong-Jun Han, Liangqi Yuan, Seyyedali Hosseinalipour, and Christopher G. Brinton. Federated sketching lora: A flexible framework for heterogeneous collaborative fine-tuning of llms, 2026. URL https://arxiv.org/abs/2501.19389.

Margalit R. Glasgow, Honglin Yuan, and Tengyu Ma. Sharp bounds for federated averaging (local SGD) and continuous perspective. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *International Conference on Artificial Intelligence and Statistics, AISTATS 2022, 28-30 March 2022, Virtual Event*, volume 151 of *Proceedings of Machine Learning Research*, pp. 9050–9090. PMLR, 2022. URL https://proceedings.mlr.press/v151/glasgow22a.html.

Yanzhao Hou, Jiaxiang Geng, Boyu Li, Xiaofeng Tao, Juncheng Wang, Xiaodong Xu, and Bing Luo. Adaptive federated lora in heterogeneous wireless networks with independent sampling. *CoRR*, abs/2505.23555, 2025. doi: 10.48550/ARXIV.2505.23555. URL https://doi.org/10.48550/arXiv.2505.23555.

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 *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

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

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

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In Inderjit S. Dhillon, Dimitris S. Papailiopoulos, and Vivienne Sze (eds.), *Proceedings of the Third Conference on Machine Learning and Systems, MLSys 2020, Austin, TX, USA, March 2-4, 2020*. mlsys.org, 2020. URL https://proceedings.mlsys.org/paper_files/paper/2020/hash/1f5fe83998a09396ebe6477d9475ba0c-Abstract.html.

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. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In Aarti Singh and Xiaojin (Jerry) Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 2017. URL http://proceedings.mlr.press/v54/mcmahan17a.html.

Siqiao Mu and Diego Klabjan. On the convergence rate of lora gradient descent. *CoRR*, abs/2512.18248, 2025. doi: 10.48550/ARXIV.2512.18248. URL https://doi.org/10.48550/arXiv.2512.18248.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 1406–1415. IEEE, 2019. doi: 10.1109/ICCV.2019.00149. URL https://doi.org/10.1109/ICCV.2019.00149.

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), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 1316–1336. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.ACL-LONG.67. URL https://doi.org/10.18653/v1/2025.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, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=NLPzL6HWNl.

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 *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=rJ4km2R5t7.

Ru Wang, Chengchang Liu, and John C. S. Lui. On the convergence of stochastic low-rank adaptation. *CoRR*, abs/2607.21975, 2026. doi: 10.48550/ARXIV.2607.21975. URL https://doi.org/10.48550/arXiv.2607.21975.

Ziyao Wang, Zheyu Shen, Yexiao He, Guoheng Sun, Hongyi Wang, Lingjuan Lyu, and Ang Li. Flora: Federated fine-tuning large language models with heterogeneous low-rank adaptations. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/28312c9491d60ed0c77f7fff4ad86dd1-Abstract-Conference.html.

Haobo Zhang, Jiankun Wang, Suraj Rajendran, Weishen Pan, Lam Tsoi, Yong Chen, Fei Wang, and Jiayu Zhou. Dysco: Dynamic subspace boosting to mitigate lora interference in federated learning. *CoRR*, abs/2607.14367, 2026a. doi: 10.48550/ARXIV.2607.14367. URL https://doi.org/10.48550/arXiv.2607.14367.

Haoran Zhang, Dongjun Kim, Seohyeon Cha, and Haris Vikalo. Fedrot-lora: Mitigating rotational misalignment in federated lora. *CoRR*, abs/2602.23638, 2026b. doi: 10.48550/ARXIV.2602.23638. URL https://doi.org/10.48550/arXiv.2602.23638.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 649–657, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/250cf8b51c773f3f8dc8b4be867a9a02-Abstract.html.

Xuechen Zhang, Mingchen Li, Xiangyu Chang, Jiasi Chen, Amit K. Roy-Chowdhury, Ananda Theertha Suresh, and Samet Oymak. Fedyolo: Augmenting federated learning with pretrained transformers, 2023. URL https://arxiv.org/abs/2307.04905.

A TECHNICAL PRELIMINARIES

We use the notation of Section 3: $X = BA$, $V = [B; A^\top]$, $J_i(V) = F_i(BA)$, and $J(V) = \sum_{i=1}^N p_i J_i(V)$. In particular, $\nabla F_i(BA)$ is the gradient with respect to the weight update X, whereas $\nabla J_i(V)$ is the stacked gradient with respect to the factors. All inner products are Frobenius. For server iterates, define

$$\Delta_t := J(V^t) - F_\star, \qquad R_t := \|V^t\|_F^2 = \|B^t\|_F^2 + \|A^t\|_F^2.$$

Replacing L by $\max\{L, 1\}$ if necessary, we take $L \geq 1$ in the bounds below.

Appendices B and C prove the FedAvg(LoRA) upper and lower bounds. Appendices D and E prove the full- and partial-participation guarantees for FedStabLoRA.

A.1 WEIGHT-SPACE GRADIENT BOUNDS

Smoothness and the lower bound on F bound $\|\nabla F(X)\|_F$. Assumption 3.3 then bounds the client gradients at the same X. No separate lower bound on each F_i is required.

Lemma A.1 (Weight-gradient bounds)**.** *Under Assumptions 3.1–3.3, for every $X \in \mathbb{R}^{m\times n}$,*

$$\|\nabla F(X)\|_F^2 \leq 2L(F(X) - F_\star), \tag{7}$$

$$\sum_i p_i \|\nabla F_i(X)\|_F^2 \leq \zeta^2 + 2L(F(X) - F_\star), \tag{8}$$

$$\max_i \|\nabla F_i(X)\|_F \leq \|\nabla F(X)\|_F + \zeta/\sqrt{p_{\min}}. \tag{9}$$

Proof. First, the weighted objective is L-smooth, since

$$\|\nabla F(X) - \nabla F(Y)\|_F \leq \sum_i p_i \|\nabla F_i(X) - \nabla F_i(Y)\|_F \leq L\|X - Y\|_F.$$

Apply its descent inequality with $Y = X - L^{-1}\nabla F(X)$:

$$\begin{aligned} F_\star \leq F(Y) &\leq F(X) + \langle \nabla F(X), Y - X\rangle + \frac{L}{2}\|Y - X\|_F^2 \\ &= F(X) - \frac{1}{2L}\|\nabla F(X)\|_F^2. \end{aligned}$$

Rearranging proves (7).

For the second assertion, write each client gradient as $\nabla F_i(X) = \nabla F(X) + [\nabla F_i(X) - \nabla F(X)]$. The cross terms in the weighted sum of squared norms vanish because $\sum_i p_i[\nabla F_i(X) - \nabla F(X)] = 0$. Thus

$$\begin{aligned} \sum_i p_i \|\nabla F_i(X)\|_F^2 &= \|\nabla F(X)\|_F^2 + \sum_i p_i \|\nabla F_i(X) - \nabla F(X)\|_F^2 \\ &\leq 2L(F(X) - F_\star) + \zeta^2, \end{aligned}$$

where the last line uses (7) and Assumption 3.3.

Finally, every summand in the heterogeneity bound is nonnegative. Hence, for each i,

$$\|\nabla F_i(X) - \nabla F(X)\|_F \leq \frac{\zeta}{\sqrt{p_i}} \leq \frac{\zeta}{\sqrt{p_{\min}}}.$$

Adding $\|\nabla F(X)\|_F$ and taking the maximum over i proves (9). □

A.2 FACTOR-SPACE GRADIENT AND SMOOTHNESS BOUNDS

We first bound the change in BA and in the factor gradients near a fixed V. The second lemma bounds gradient differences between any two points in a larger ball; it will be used for consecutive server iterates in the FedStabLoRA proof.

Lemma A.2 (Product and local gradient bounds). *Let $V = [B; A^\top]$ and $U = [\widetilde{B}; \widetilde{A}^\top]$. Then*

$$\|BA\|_F \le \tfrac{1}{2}\|V\|_F^2, \qquad \|\widetilde{B}\widetilde{A} - BA\|_F \le \tfrac{1}{2}(\|U\|_F + \|V\|_F)\|U - V\|_F. \tag{10}$$

Under Assumption 3.1, $\|\nabla J_i(V)\|_F \le \|V\|_F \, \|\nabla F_i(BA)\|_F$. If $\|U - V\|_F \le \|V\|_F$, then

$$\|\nabla J_i(U) - \nabla J_i(V)\|_F \le \big(\|\nabla F_i(BA)\|_F + 3L\|V\|_F^2\big)\|U - V\|_F. \tag{11}$$

Define $\Lambda(V) := 1 + \max_i \|\nabla F_i(BA)\|_F + 4L\|V\|_F^2$. For every factor displacement Z with $\|Z\|_F \le \|V\|_F$,

$$J(V + Z) \le J(V) + \langle \nabla J(V), Z\rangle + \tfrac{1}{2}\Lambda(V)\|Z\|_F^2. \tag{12}$$

Proof. **Products of the factors.** Submultiplicativity of the Frobenius norm and $2ab \le a^2 + b^2$ give $\|BA\|_F \le \|B\|_F\|A\|_F \le \|V\|_F^2/2$. For the product difference, expand

$$\widetilde{B}\widetilde{A} - BA = \tfrac{1}{2}[(\widetilde{B} - B)(\widetilde{A} + A) + (\widetilde{B} + B)(\widetilde{A} - A)].$$

The triangle inequality, followed by scalar Cauchy–Schwarz, yields

$$\begin{aligned}\|\widetilde{B}\widetilde{A} - BA\|_F &\le \tfrac{1}{2}\big(\|\widetilde{B} - B\|_F\|\widetilde{A} + A\|_F + \|\widetilde{A} - A\|_F\|\widetilde{B} + B\|_F\big)\\ &\le \tfrac{1}{2}\|U - V\|_F\|U + V\|_F\\ &\le \tfrac{1}{2}(\|U\|_F + \|V\|_F)\|U - V\|_F.\end{aligned}$$

Factor gradients. The formulas in Section 3 imply

$$\begin{aligned}\|\nabla J_i(V)\|_F^2 &= \|\nabla F_i(BA)A^\top\|_F^2 + \|B^\top \nabla F_i(BA)\|_F^2\\ &\le (\|A\|_F^2 + \|B\|_F^2)\|\nabla F_i(BA)\|_F^2.\end{aligned}$$

To compare gradients at U and V, expand their two blocks as

$$\begin{aligned}\nabla_B J_i(U) - \nabla_B J_i(V) &= [\nabla F_i(\widetilde{B}\widetilde{A}) - \nabla F_i(BA)]\widetilde{A}^\top + \nabla F_i(BA)(\widetilde{A} - A)^\top,\\ \nabla_A J_i(U) - \nabla_A J_i(V) &= \widetilde{B}^\top[\nabla F_i(\widetilde{B}\widetilde{A}) - \nabla F_i(BA)] + (\widetilde{B} - B)^\top \nabla F_i(BA).\end{aligned}$$

Take the norm after stacking the B block and the transpose of the A block. Bound the terms containing the gradient difference together, and then the terms containing the factor differences together. This gives

$$\begin{aligned}\|\nabla J_i(U) - \nabla J_i(V)\|_F &\le \|U\|_F\|\nabla F_i(\widetilde{B}\widetilde{A}) - \nabla F_i(BA)\|_F + \|\nabla F_i(BA)\|_F\|U - V\|_F\\ &\le \big[\|\nabla F_i(BA)\|_F + \tfrac{L}{2}\|U\|_F(\|U\|_F + \|V\|_F)\big]\,\|U - V\|_F.\end{aligned}$$

The second line uses Assumption 3.1 and (10). If $\|U - V\|_F \le \|V\|_F$, then $\|U\|_F \le 2\|V\|_F$ and the second term in brackets is at most $3L\|V\|_F^2$. This proves (11).

Objective descent. For $\tau \in [0, 1]$, the point $V + \tau Z$ satisfies the same neighborhood condition. Averaging (11) over clients gives $\|\nabla J(V + \tau Z) - \nabla J(V)\|_F \le \Lambda(V)\tau\|Z\|_F$. Integration along the segment now gives

$$\begin{aligned}J(V + Z) - J(V) - \langle \nabla J(V), Z\rangle &= \int_0^1 \langle \nabla J(V + \tau Z) - \nabla J(V), Z\rangle \, d\tau\\ &\le \Lambda(V)\|Z\|_F^2 \int_0^1 \tau \, d\tau = \tfrac{1}{2}\Lambda(V)\|Z\|_F^2.\end{aligned}$$

When $V = 0$, the neighborhood condition forces $U = V$ and $Z = 0$, so the local assertions hold directly. □

Lemma A.3 (Smoothness on a ball of factors). *Under Assumption 3.1, on $\mathcal{B}(V) := \{U : \|U - V\|_F \le 1 + \|V\|_F\}$, each ∇J_i and ∇J is Lipschitz with constant*

$$\ell(V) := \max_i \|\nabla F_i(BA)\|_F + 16L(1 + \|V\|_F^2).$$

The descent inequality with constant $\ell(V)$ therefore holds between any two points in this ball.

Proof. Every $U = [\widetilde{B}; \widetilde{A}^\top] \in \mathcal{B}(V)$ satisfies $\|U\|_F \le 1 + 2\|V\|_F$. Therefore (10) gives

$$\|\widetilde{B}\widetilde{A} - BA\|_F \le \tfrac{1}{2}(1 + 3\|V\|_F)(1 + \|V\|_F) \le \tfrac{5}{2}(1 + \|V\|_F^2).$$

For the last inequality, use $2\|V\|_F \le 1 + \|V\|_F^2$. Weight-gradient smoothness then gives

$$\|\nabla F_i(\widetilde{B}\widetilde{A})\|_F \le \|\nabla F_i(BA)\|_F + \tfrac{5}{2}L(1 + \|V\|_F^2).$$

Now take any $U, W \in \mathcal{B}(V)$. The general gradient-difference estimate derived in the preceding proof, with W as its reference point, has coefficient at most

$$\begin{aligned}&\|\nabla F_i(BA)\|_F + \tfrac{5}{2}L(1 + \|V\|_F^2) + L(1 + 2\|V\|_F)^2 \\ &\quad \le \|\nabla F_i(BA)\|_F + 16L(1 + \|V\|_F^2),\end{aligned}$$

because $(1 + 2\|V\|_F)^2 \le 8(1 + \|V\|_F^2)$. This bound applies to arbitrary pairs U, W, not only to pairs with one point equal to the center. Averaging with the weights p_i gives the same coefficient for ∇J. Finally, the ball is convex, so the segment from U to W stays in it. Integrating the gradient difference along that segment proves the stated descent inequality. □

B Proof of the FedAvg(LoRA) Upper Bound

We prove Theorem 4.1 with a local stepsize that is fixed across rounds. We first bound the local displacements and the error in the server update. A variance identity bounds this error by a global gradient term and a term proportional to ζ^2. Combining the resulting descent inequality with a bound on factor movement gives the convergence rate.

B.1 Local Trajectory and Drift Bounds

Let $\gamma := K\eta$. The client recursion in Section 3 first gives

$$V_{i,K}^t = V^t - \eta \sum_{e=0}^{K-1} \nabla J_i(V_{i,e}^t).$$

Average this identity with weights p_i, and add and subtract $\eta \sum_{e=0}^{K-1} \sum_i p_i \nabla J_i(V^t) = \gamma \nabla J(V^t)$. The resulting server update is

$$\begin{aligned}V^{t+1} &= V^t - \gamma\big(\nabla J(V^t) + d_t\big), \\ d_t &:= \frac{1}{K} \sum_{e=0}^{K-1} \sum_{i=1}^{N} p_i \big[\nabla J_i(V_{i,e}^t) - \nabla J_i(V^t)\big].\end{aligned} \tag{13}$$

The error d_t measures the change in the client gradients during a round and vanishes when $K = 1$. Write $\Lambda_t := \Lambda(V^t)$, with Λ as defined in Lemma A.2.

Lemma B.1 (One FedAvg round)**.** *If $\gamma\Lambda_t \le 1/8$, then*

$$\|V_{i,e}^t - V^t\|_F \le 2\gamma \|\nabla J_i(V^t)\|_F \le \tfrac{1}{4}\sqrt{R_t}, \qquad 0 \le e \le K, \tag{14}$$

$$\|d_t\|_F^2 \le 4\gamma^2 \Lambda_t^2 R_t \sum_i p_i \|\nabla F_i(B^t A^t)\|_F^2, \tag{15}$$

$$\Delta_{t+1} \le \Delta_t - \tfrac{\gamma}{2}\|\nabla J(V^t)\|_F^2 + \tfrac{\gamma}{2}\|d_t\|_F^2. \tag{16}$$

Proof. If $R_t = 0$, then $A^t = B^t = 0$, all factor gradients at the starting point vanish, and every local iterate and the server iterate stay at zero. Suppose henceforth that $R_t > 0$.

Step 1: bound the local iterates. Fix client i. Provided the earlier iterates lie within distance $\|V^t\|_F$ of V^t, the local update and Lemma A.2 give

$$\begin{aligned}\|V_{i,e+1}^t - V^t\|_F &\le \|V_{i,e}^t - V^t\|_F + \eta \|\nabla J_i(V_{i,e}^t)\|_F \\ &\le (1 + \eta\Lambda_t)\|V_{i,e}^t - V^t\|_F + \eta \|\nabla J_i(V^t)\|_F.\end{aligned}$$

Starting from $V_{i,0}^t = V^t$, iteration of this inequality yields

$$\begin{aligned}\|V_{i,e}^t - V^t\|_F &\le \eta \sum_{k=0}^{e-1} (1+\eta\Lambda_t)^k \|\nabla J_i(V^t)\|_F \\ &\le e\eta \exp(e\eta\Lambda_t)\|\nabla J_i(V^t)\|_F \\ &\le 2e\eta\|\nabla J_i(V^t)\|_F.\end{aligned}$$

Here $e\eta\Lambda_t \le K\eta\Lambda_t \le 1/8$, so $\exp(e\eta\Lambda_t) < 2$. Also, the factor-gradient bound and the definition of Λ_t imply $\|\nabla J_i(V^t)\|_F \le \Lambda_t\sqrt{R_t}$. Consequently the next iterate is within $2\gamma\Lambda_t\sqrt{R_t} \le \sqrt{R_t}/4$ of V^t. The starting iterate satisfies the neighborhood condition, and this bound preserves it at each step. This proves the induction for $e = 0, \dots, K$, rather than assuming that the trajectory stays in the ball.

Step 2: bound the server-update error. The weights p_i/K in the definition of d_t sum to one. Jensen's inequality, the local gradient bound, and Step 1 give

$$\begin{aligned}\|d_t\|_F^2 &\le \frac{1}{K}\sum_{e=0}^{K-1}\sum_i p_i \|\nabla J_i(V_{i,e}^t) - \nabla J_i(V^t)\|_F^2 \\ &\le \frac{\Lambda_t^2}{K}\sum_{e=0}^{K-1}\sum_i p_i \|V_{i,e}^t - V^t\|_F^2 \\ &\le 4\gamma^2\Lambda_t^2 \sum_i p_i \|\nabla J_i(V^t)\|_F^2 \\ &\le 4\gamma^2\Lambda_t^2 R_t \sum_i p_i \|\nabla F_i(B^tA^t)\|_F^2.\end{aligned}$$

Step 3: apply descent at the server iterate. Since $V^{t+1} - V^t = \sum_i p_i(V_{i,K}^t - V^t)$, Step 1 also gives $\|V^{t+1} - V^t\|_F \le \sqrt{R_t}/4$. We may therefore apply (12) with this displacement and substitute (13):

$$\begin{aligned}\Delta_{t+1} - \Delta_t &\le -\gamma\langle \nabla J(V^t), \nabla J(V^t) + d_t\rangle + \frac{\Lambda_t\gamma^2}{2}\|\nabla J(V^t) + d_t\|_F^2 \\ &\le -\gamma\langle \nabla J(V^t), \nabla J(V^t) + d_t\rangle + \frac{\gamma}{2}\|\nabla J(V^t) + d_t\|_F^2 \\ &= -\frac{\gamma}{2}\|\nabla J(V^t)\|_F^2 + \frac{\gamma}{2}\|d_t\|_F^2.\end{aligned}$$

The second line uses $\gamma\Lambda_t \le 1$; expanding the squared norm in that line cancels the cross term and gives the last equality. □

B.2 Client Heterogeneity in the Descent Bound

At a common factor point V, the matrices A and B are the same for all clients. The variance identity therefore separates $\|\nabla J(V)\|_F^2$ from a term bounded by $\|V\|_F^2\zeta^2$. We keep the global gradient term because it can be combined with objective descent.

Lemma B.2 (Factor-gradient variance bound)**.** *Under Assumption 3.3, at every factor point V,*

$$\sum_i p_i \|\nabla J_i(V)\|_F^2 \le \|\nabla J(V)\|_F^2 + \|V\|_F^2\zeta^2. \tag{17}$$

Proof. The chain rule from Section 3 gives

$$\begin{aligned}\|\nabla J_i(V) - \nabla J(V)\|_F^2 &= \|[\nabla F_i(BA) - \nabla F(BA)]A^\top\|_F^2 \\ &\quad + \|B^\top[\nabla F_i(BA) - \nabla F(BA)]\|_F^2 \\ &\le \|V\|_F^2\|\nabla F_i(BA) - \nabla F(BA)\|_F^2.\end{aligned}$$

Using $\sum_i p_i \nabla J_i(V) = \nabla J(V)$, the weighted variance identity and Assumption 3.3 yield

$$\begin{aligned}\sum_i p_i \|\nabla J_i(V)\|_F^2 &= \|\nabla J(V)\|_F^2 + \sum_i p_i \|\nabla J_i(V) - \nabla J(V)\|_F^2 \\ &\le \|\nabla J(V)\|_F^2 + \|V\|_F^2 \zeta^2.\end{aligned}$$

□

B.3 One-Round Descent

Let $\theta_K := (K-1)/K$. Combining the local-displacement and variance bounds gives a drift bound with one term proportional to $\|\nabla J(V^t)\|_F^2$ and another proportional to $\theta_K^2 \zeta^2$. The first term is small enough to retain a decrease in the objective.

Lemma B.3 (One-round descent with client heterogeneity)**.** *If $\gamma \Lambda_t \le 1/8$, then*

$$\begin{aligned}\|d_t\|_F^2 &\le 4\theta_K^2 \gamma^2 \Lambda_t^2 (\|\nabla J(V^t)\|_F^2 + R_t \zeta^2) \\ &\le \tfrac{1}{16} \|\nabla J(V^t)\|_F^2 + 4\gamma^2 \Lambda_t^2 R_t \theta_K^2 \zeta^2, \qquad (18) \\ \Delta_{t+1} &\le \Delta_t - \tfrac{\gamma}{4} \|\nabla J(V^t)\|_F^2 + 2\gamma^3 \Lambda_t^2 R_t \theta_K^2 \zeta^2. \qquad (19)\end{aligned}$$

Proof. Lemma B.1 already verifies the neighborhood condition for every local iterate. Its proof gives the sharper bound $\|V_{i,e}^t - V^t\|_F \le 2e\eta \|\nabla J_i(V^t)\|_F$. Only $e = 0, \dots, K-1$ occur in the gradient evaluations, so $e\eta \le (K-1)\eta = \theta_K \gamma$. Using these bounds in the Jensen step gives

$$\begin{aligned}\|d_t\|_F^2 &\le \frac{\Lambda_t^2}{K} \sum_{e=0}^{K-1} \sum_i p_i \|V_{i,e}^t - V^t\|_F^2 \\ &\le 4\theta_K^2 \gamma^2 \Lambda_t^2 \sum_i p_i \|\nabla J_i(V^t)\|_F^2 \\ &\le 4\theta_K^2 \gamma^2 \Lambda_t^2 (\|\nabla J(V^t)\|_F^2 + R_t \zeta^2).\end{aligned}$$

The last line is Lemma B.2. Since $0 \le \theta_K \le 1$ and $\gamma \Lambda_t \le 1/8$, $4\theta_K^2 \gamma^2 \Lambda_t^2 \le 1/16$, proving (18).

Substitute that inequality into (16):

$$\begin{aligned}\Delta_{t+1} &\le \Delta_t - \frac{\gamma}{2} \|\nabla J(V^t)\|_F^2 + \frac{\gamma}{2} \left[\frac{1}{16} \|\nabla J(V^t)\|_F^2 + 4\gamma^2 \Lambda_t^2 R_t \theta_K^2 \zeta^2 \right] \\ &= \Delta_t - \frac{15\gamma}{32} \|\nabla J(V^t)\|_F^2 + 2\gamma^3 \Lambda_t^2 R_t \theta_K^2 \zeta^2 \\ &\le \Delta_t - \frac{\gamma}{4} \|\nabla J(V^t)\|_F^2 + 2\gamma^3 \Lambda_t^2 R_t \theta_K^2 \zeta^2.\end{aligned}$$

For $K = 1$, the same argument has $\theta_K = 0$ and $d_t = 0$. □

B.4 Objective and Factor-Norm Control

Fix a finite upper bound $\bar{\zeta}$ on ζ. The following constants specify one admissible choice of the prefactor a_0 in Theorem 4.1:

$$\begin{aligned}R_* &= 2R_0 + 24(\Delta_0 + 1) + 1, \\ \Lambda_* &= 1 + \sqrt{2L(\Delta_0 + 1)} + \bar{\zeta}/\sqrt{p_{\min}} + 4LR_*, \\ a_0 &= \min\left\{1, \frac{1}{16\Lambda_*}, [32\Lambda_*^2 R_* (1 + \bar{\zeta})]^{-1/3}\right\}.\end{aligned} \qquad (20)$$

Fix $0 < a \le a_0$ and set

$$\gamma := K\eta = \frac{a}{\sqrt{T} + (\theta_K \zeta)^{1/3} T^{2/3}}, \qquad (21)$$

which is the local stepsize in (1) multiplied by K.

Lemma B.4 (Objective and factor-norm bounds)**.** *Under the assumptions of Theorem 4.1 and the choices above, for every* $0 \le t \le T$*,*

$$\Delta_t < \Delta_0 + 1, \qquad R_t < R_*(1 + T\gamma). \tag{22}$$

Moreover, for every $0 \le k \le T$*,* $\sum_{t<k} \gamma \|\nabla J(V^t)\|_F^2 \le 4\Delta_0 + 1$*.*

Proof. **Step 1: verify descent up to a possible first exit.** Both bounds in (22) hold at $t = 0$ because $R_* > R_0$ and $\gamma > 0$. Suppose $k \in \{1, \dots, T\}$ is the first index at which one of them fails. For each $t < k$, the two bounds at that round and Lemma A.1 imply

$$\begin{aligned}\max_i \|\nabla F_i(B^t A^t)\|_F &\le \sqrt{2L(\Delta_0 + 1)} + \frac{\bar{\zeta}}{\sqrt{p_{\min}}}, \\ \Lambda_t &\le 1 + \sqrt{2L(\Delta_0 + 1)} + \frac{\bar{\zeta}}{\sqrt{p_{\min}}} + 4LR_*(1 + T\gamma) \\ &\le \Lambda_*(1 + T\gamma).\end{aligned}$$

The stepsize satisfies $\gamma \le a/\sqrt{T}$. Thus

$$\gamma\Lambda_t \le \Lambda_*(\gamma + T\gamma^2) \le \Lambda_*(a + a^2) \le 2\Lambda_* a \le \tfrac{1}{8},$$

where we used $T \ge 1$, $a \le 1$, and $a \le (16\Lambda_*)^{-1}$. Consequently Lemma B.3 applies to every round $t < k$. No property of the possibly failing iterate V^k has been used.

Step 2: sum the heterogeneity error. Define $P_k := \sum_{t<k} 4\gamma^3 \Lambda_t^2 R_t \theta_K^2 \zeta^2$. Step 1, the factor bound for $t < k$, and $k \le T$ give

$$\begin{aligned}P_k &\le 4\Lambda_*^2 R_* (\theta_K \zeta)^2 T\gamma^3 (1 + T\gamma)^3 \\ &\le 16\Lambda_*^2 R_* \big[(\theta_K \zeta)^2 T\gamma^3 + (\theta_K \zeta)^2 T^4 \gamma^6\big].\end{aligned}$$

The second line uses $(1 + x)^3 \le 4(1 + x^3)$ for $x \ge 0$. If $\theta_K \zeta = 0$, then $P_k = 0$. Otherwise the second term in the stepsize denominator implies $\gamma \le a(\theta_K \zeta)^{-1/3} T^{-2/3}$, and hence

$$(\theta_K \zeta)^2 T\gamma^3 \le \frac{a^3 \theta_K \zeta}{T}, \qquad (\theta_K \zeta)^2 T^4 \gamma^6 \le a^6.$$

Using $\theta_K \zeta \le \bar{\zeta}$, $T \ge 1$, $a^6 \le a^3$, and (20), we obtain

$$P_k \le 16\Lambda_*^2 R_* a^3 (1 + \bar{\zeta}) \le \tfrac{1}{2}.$$

Step 3: bound the objective and the gradient sum. Sum (19) over $t = 0, \dots, k - 1$. The objective terms telescope, and the sum of the positive error terms is $P_k/2$. Therefore

$$\Delta_k + \tfrac{1}{4} \sum_{t<k} \gamma \|\nabla J(V^t)\|_F^2 \le \Delta_0 + \tfrac{1}{2} P_k. \tag{23}$$

Dropping the gradient sum shows $\Delta_k \le \Delta_0 + 1/4 < \Delta_0 + 1$. Alternatively, using $\Delta_k \ge 0$ gives

$$\sum_{t<k} \gamma \|\nabla J(V^t)\|_F^2 \le 4\Delta_0 + 2P_k.$$

Step 4: use the gradient sum to bound factor movement. Sum (18) after multiplying it by γ:

$$\sum_{t<k} \gamma \|d_t\|_F^2 \le \tfrac{1}{16} \sum_{t<k} \gamma \|\nabla J(V^t)\|_F^2 + P_k.$$

Applying $\|U + W\|_F^2 \le 2\|U\|_F^2 + 2\|W\|_F^2$ then gives

$$\begin{aligned}\sum_{t<k} \gamma \|\nabla J(V^t) + d_t\|_F^2 &\le \tfrac{17}{8} \sum_{t<k} \gamma \|\nabla J(V^t)\|_F^2 + 2P_k \\ &\le \tfrac{17}{2}\Delta_0 + \tfrac{25}{4} P_k < 12(\Delta_0 + 1).\end{aligned}$$

Now sum the exact server updates. Weighted Cauchy–Schwarz yields

$$\begin{aligned}\|V^k - V^0\|_F &\le \sum_{t<k} \gamma \|\nabla J(V^t) + d_t\|_F \\ &\le \left(\sum_{t<k} \gamma\right)^{1/2} \left(\sum_{t<k} \gamma \|\nabla J(V^t) + d_t\|_F^2\right)^{1/2} \\ &\le \sqrt{12(\Delta_0 + 1)k\gamma}.\end{aligned}$$

Consequently,

$$\begin{aligned}R_k &\le 2R_0 + 2\|V^k - V^0\|_F^2 \\ &\le 2R_0 + 24(\Delta_0 + 1)k\gamma < R_*(1 + T\gamma).\end{aligned}$$

Both proposed exit conditions fail, a contradiction. The two bounds therefore hold through round T. Repeating Steps 2–3 for any $k \le T$ gives $P_k \le 1/2$ and $\sum_{t<k} \gamma \|\nabla J(V^t)\|_F^2 \le 4\Delta_0 + 1$. Thus the argument derives bounded factor norms on the chosen horizon; it does not assume them in order to apply descent. □

B.5 Convergence Rate and Special Cases

Proof of Theorem 4.1. Choose a_0 as in (20). Lemma B.4 gives $\sum_{t<T} \gamma \|\nabla J(V^t)\|_F^2 \le 4\Delta_0 + 1$. Dividing by $T\gamma$ gives

$$\frac{1}{T} \sum_{t<T} \|\nabla J(V^t)\|_F^2 \le \frac{4\Delta_0 + 1}{T\gamma} = \frac{4\Delta_0 + 1}{a} \left(T^{-1/2} + (\theta_K \zeta)^{1/3} T^{-1/3}\right). \tag{24}$$

Since $\theta_K = 1 - 1/K$, this proves (2). In particular,

$$T \ge \left\lceil \max\left\{1, \frac{4(4\Delta_0 + 1)^2}{a^2 \epsilon^4}, \frac{8(4\Delta_0 + 1)^3 \theta_K \zeta}{a^3 \epsilon^6}\right\} \right\rceil$$

bounds each of the two terms on the right of (24) by $\epsilon^2/2$. When $\theta_K \zeta = 0$, the second term is identically zero and imposes no additional restriction. Since

$$\min_{0 \le t < T} \|\nabla J(V^t)\|_F^2 \le \frac{1}{T} \sum_{t<T} \|\nabla J(V^t)\|_F^2 \le \epsilon^2,$$

at least one server iterate is ϵ-stationary. The maximum of the two horizon terms is no greater than their sum, giving the best-iterate communication complexity

$$O\left(\epsilon^{-4} + \theta_K \zeta\, \epsilon^{-6}\right). \tag{25}$$

For fixed a, the hidden constants are independent of K and of $\zeta \in [0, \bar{\zeta}]$. □

Special cases. When $K = 1$ or $\zeta = 0$, (25) reduces to $O(\epsilon^{-4})$. For $\zeta = 0$ and $K > 1$, d_t need not vanish: clients still evaluate gradients away from the round-start point. In that case its contribution is absorbed into descent by (18). The heterogeneity-dependent communication term is linear in ζ, although Assumption 3.3 bounds a squared difference by ζ^2. The lower bound matches the ϵ^{-6} term in the worst case for fixed $K \ge 2$.

For a single problem, one may set $\bar{\zeta} = \zeta$. For a family of problems with decreasing heterogeneity, fixing $\bar{\zeta}$ once ensures that the constants remain bounded as $\zeta \to 0$.

C Proof of the Matching Lower Bound

We prove Theorem 4.2 by constructing two client objectives for each fixed $K \ge 2$ and sufficiently small ϵ. The objectives do not depend on the stepsize sequence. We first verify the assumptions, derive the local updates and their server average, and then bound progress through a collection of slow intervals.

Write $g = \Theta_K(\epsilon^2)$ for the scale specified below. After removing a short segment from the left end of each slow interval, the remaining intervals have total length $\Omega_K(g^{-1})$. A forward round starting outside the slow intervals cannot enter these remaining segments, and a round starting inside a slow interval advances by at most $O_K(g^2)$. These bounds give $\Omega_K(g^{-3})$ communication rounds without requiring the server coordinate to increase monotonically.

C.1 Hard Instance and Assumption Verification

Let e_1, e_2 be the standard basis of $\mathbb{R}^2$ and define

$$E_x = e_1 e_1^\top, \qquad E_y = \frac{e_1 e_2^\top + e_2 e_1^\top}{\sqrt{2}}, \qquad x(X) = \langle E_x, X\rangle, \quad y(X) = \langle E_y, X\rangle.$$

The matrices E_x, E_y are orthonormal in the Frobenius inner product. The coordinate $x(X) = X_{11}$ measures progress, while $y(X)$ measures the transverse displacement caused by the client-specific terms. We label the two clients by $\sigma \in \{+1, -1\}$ and set $p_+ = p_- = 1/2$, so $F = (F_+ + F_-)/2$. We use rank-one factors $B \in \mathbb{R}^{2\times 1}$ and $A \in \mathbb{R}^{1\times 2}$, with the FedAvg(LoRA) update from Section 3. Thus every round uses one stepsize $\eta_t \in (0, 1]$ across both clients, both factors, and all K local steps, followed by separate averaging of A and B.

Theorem C.1 (An explicit hard instance)**.** *For every number of local steps $K \geq 2$ and every $0 < \epsilon \leq \epsilon_0(K) := (222200000K^2)^{-1}$, there are globally $C^{1,1}$, 1-smooth client objectives such that*

$$\frac{1}{2}\sum_{\sigma=\pm 1} \|\nabla F_\sigma(X) - \nabla F(X)\|_F^2 \leq \frac{1}{40000^2} \qquad \textit{for all } X.$$

The global objective attains its minimum at $X_\star = -E_x$, with factors $B_\star = -e_1$ and $A_\star = e_1^\top$. Take $F_\star = \min_X F(X)$ for this instance. At the initialization $B^0 = 0$, $A^0 = e_1^\top$,

$$R_0 = 1, \qquad \Delta_0 < 1/10, \qquad \|X_\star - B^0 A^0\|_F = 1,$$
$$\|B^0 - B_\star\|_F^2 + \|A^0 - A_\star\|_F^2 = 1.$$

On this same instance, every sequence $(\eta_t)_{t\geq 0} \in (0, 1]^{\mathbb{N}_0}$ satisfies

$$\|\nabla J(V^t)\|_F > \epsilon \qquad \textit{for every round } 0 \leq t < \frac{74}{(22220K)^6}\epsilon^{-6}. \tag{26}$$

The displayed bounds on smoothness, heterogeneity, initialization, and distances do not depend on K or ϵ. The communication-bound coefficient may depend on K.

For each fixed (K, ϵ), this is one instance on which the bound holds for every stepsize sequence in $(0, 1]$.

Client objectives. Fix $K \geq 2$ and define the constants below. They depend only on K; the scale g will also depend on ϵ. The slow intervals have width w and are extended by w_{g} on each side. The constant J_0 bounds the increase in x near these intervals, and R_y is the threshold beyond which ψ is constant. The subscript h distinguishes construction constants from the FedStabLoRA parameters.

$$h = \frac{1}{10}, \quad M_{\mathrm{h}} = 10^6 K^2, \quad \mu_{\mathrm{h}} = {M_{\mathrm{h}}}^{-1/2} = \frac{1}{1000K}, \quad \kappa_K = \frac{1}{22000K}, \tag{27}$$

$$R_y = 48K + 2, \quad U = 40R_y, \quad J_0 = KU, \quad w = 10J_0, \quad w_{\mathrm{g}} = 2J_0, \tag{28}$$

$$\ell_{\mathrm{c}} = w + 2w_{\mathrm{g}} = 14J_0, \quad P = 2\ell_{\mathrm{c}} + 2, \quad \lambda_{\mathrm{h}} = \frac{1/20}{\ell_{\mathrm{c}} + 1/2},$$
$$\Gamma = \frac{10K\mu_{\mathrm{h}}}{\lambda_{\mathrm{h}}}, \qquad \nu = \frac{1}{40K}. \tag{29}$$

For the specified accuracy, set

$$g = (22220K\epsilon)^2 = \left(\frac{101\epsilon}{100\kappa_K}\right)^2, \qquad g_0 = \frac{1}{100M_{\mathrm{h}}}, \qquad K_{\mathrm{c}} = \lfloor g^{-1}\rfloor,$$
$$x_c = M_{\mathrm{h}}/g, \qquad S = x_c + K_{\mathrm{c}}P, \qquad R_f = S + 2.$$

The accuracy restriction gives $g \leq (22220K/(222200000K^2))^2 = 1/(10^8K^2) = g_0$. In particular, $\sqrt{g} \leq 1/(10000K)$ and $g_0 < 1$. The constants satisfy

$$1920K^2 \leq J_0 \leq 1960K^2, \qquad \lambda_{\mathrm{h}} < \frac{2\cdot 10^{-6}}{K^2}, \tag{30}$$

$$\Gamma = 2.8J_0 + 0.1 \leq 5500K^2, \qquad \Gamma g \leq 5.5\cdot 10^{-5}, \tag{31}$$

$$K(\sqrt{g} + \tfrac{3}{2}\lambda_{\mathrm{h}}) < 1.02\cdot 10^{-4}, \qquad K\xi \leq \frac{1}{22000} \quad (0 < \xi \leq \kappa_K). \tag{32}$$

For example, $J_0 = 1920K^2 + 80K \le 1960K^2$ because $K \ge 2$. Also $\lambda_{\mathrm{h}} = 1/[20(14J_0 + 1/2)]$ and $\Gamma = 0.2(14J_0 + 1/2) = 2.8J_0 + 0.1$, proving the first two lines of the bounds above. Combining $g \le 1/(10^8K^2)$ with $\Gamma \le 5500K^2$ proves $\Gamma g \le 5.5 \cdot 10^{-5}$. Finally, $K(\sqrt{g} + 3\lambda_{\mathrm{h}}/2) < 10^{-4} + 3 \cdot 10^{-6}/K \le 1.015 \cdot 10^{-4}$, which implies (32).

Transverse term and common penalty. For $y \ge 0$, define

$$\psi(y) = \begin{cases} y, & 0 \le y \le 1, \\ 1 + \displaystyle\int_1^y (2 - z)\, dz, & 1 \le y \le 2, \\ \dfrac{3}{2} - \omega \displaystyle\int_2^y \sin^2\!\left(\dfrac{\pi(z-2)}{48K}\right) dz, & 2 \le y \le R_y, \\ \Gamma g, & y \ge R_y, \end{cases} \qquad \omega = \frac{3 - 2\Gamma g}{48K},$$

and for $y < 0$, set $\psi(y) = -\psi(-y)$. To check the joins, note that $\int_1^2 (2 - z)\, dz = 1/2$ and $\int_2^{R_y} \sin^2(\pi(z-2)/(48K))\, dz = 24K$. The third piece therefore ends at $3/2 - 24K\omega = \Gamma g$, as required by the last piece. For positive y away from the joins, the derivative is

$$\psi'(y) = \begin{cases} 1, & 0 < y < 1, \\ 2 - y, & 1 < y < 2, \\ -\omega \sin^2(\pi(y-2)/(48K)), & 2 < y < R_y, \\ 0, & y > R_y. \end{cases}$$

These derivatives match at $1, 2, R_y$; the odd extension is C^1 at zero as well. Since $0 < \omega \le 1/(16K)$, the third piece is nonincreasing from $3/2$ to $\Gamma g > 0$, and its derivative lies in $[-1/(16K), 0]$. Its second derivative has magnitude at most $\omega\pi/(48K) < 1$; on $(1, 2)$ the second derivative is -1, and on the other pieces it is zero. Continuity of ψ' at the joins therefore gives

$$|\psi| \le \frac{3}{2}, \quad |\psi'| \le 1, \quad \mathrm{Lip}(\psi') \le 1, \quad \psi' \ge -\frac{1}{16K}.$$

On the positive half-line, $\psi \ge 0$ and

$$d \ge \Gamma g \implies \psi(d) \ge \Gamma g, \tag{33}$$

$$d \ge R_y \implies (\psi(d), \psi'(d)) = (\Gamma g, 0), \tag{34}$$

$$d \ge 2 \implies \psi'(d) \le 0. \tag{35}$$

The common penalty is $\nu\phi_{\mathrm{H}}(y)$, where

$$\phi_{\mathrm{H}}(y) = \begin{cases} y^2/2, & |y| \le 1, \\ |y| - 1/2, & |y| \ge 1. \end{cases}$$

Thus $\phi_{\mathrm{H}}'(y) = \max(-1, \min(y, 1))$.

Amplitude and slow intervals. Let v be the P-periodic continuous extension of

$$v(t) = \begin{cases} -\lambda_{\mathrm{h}}, & 0 \le t \le \ell_{\mathrm{c}}, \\ -\lambda_{\mathrm{h}} + 2\lambda_{\mathrm{h}}(t - \ell_{\mathrm{c}}), & \ell_{\mathrm{c}} \le t \le \ell_{\mathrm{c}} + 1, \\ \lambda_{\mathrm{h}}, & \ell_{\mathrm{c}} + 1 \le t \le 2\ell_{\mathrm{c}} + 1, \\ \lambda_{\mathrm{h}} - 2\lambda_{\mathrm{h}}(t - 2\ell_{\mathrm{c}} - 1), & 2\ell_{\mathrm{c}} + 1 \le t \le 2\ell_{\mathrm{c}} + 2. \end{cases}$$

Set $\widehat{H}(t) = h + \lambda_{\mathrm{h}}(\ell_{\mathrm{c}} + 1/4) + \int_0^t v(u)\, du$ and

$$H(x) = \begin{cases} h, & x \le x_c, \\ \widehat{H}(\ell_{\mathrm{c}} + 1/2 + x - x_c), & x_c \le x \le S, \\ h, & x \ge S. \end{cases}$$

The integral of v over one period is zero. On $[0, P]$, v is negative up to $\ell_{\mathrm{c}} + 1/2$ and positive from there to $2\ell_{\mathrm{c}} + 3/2$. Moreover,

$$\int_0^{\ell_{\mathrm{c}} + 1/2} v(u)\, du = -\lambda_{\mathrm{h}}(\ell_{\mathrm{c}} + 1/4).$$

Thus $\widehat{H}$ has minimum h at phase $\ell_{\mathrm{c}}+1/2$ and maximum $h+\lambda_{\mathrm{h}}(\ell_{\mathrm{c}}+1/2)=3/20$ half a period later. Since $S-x_c=K_{\mathrm{c}}P$, the middle branch of H starts and ends at that same minimum phase, with derivative zero at both joins. The piecewise slopes of v have magnitude at most $2\lambda_{\mathrm{h}}$. Consequently

$$h \le H \le \frac{3}{20} < 2h, \qquad |H'| \le \lambda_{\mathrm{h}}, \qquad \mathrm{Lip}(H') \le 2\lambda_{\mathrm{h}}.$$

For $j=1,\dots,K_{\mathrm{c}}$, put

$$\ell_j = x_c + \ell_{\mathrm{c}} + \frac{3}{2} + w_{\mathrm{g}} + (j-1)P, \quad I_j=[\ell_j,\ell_j+w], \quad W_j=[\ell_j-w_{\mathrm{g}},\ell_j+w+w_{\mathrm{g}}].$$

Then

$$H' = -\lambda_{\mathrm{h}} \text{ on } W_j, \qquad \ell_{K_{\mathrm{c}}}+w+w_{\mathrm{g}} = S-\frac{1}{2}, \qquad Z := \ell_{K_{\mathrm{c}}}+w < R_f.$$

Indeed, at the left endpoint of W_j, the argument of $\widehat{H}$ is

$$\ell_{\mathrm{c}} + \tfrac{1}{2} + (\ell_j - w_{\mathrm{g}}) - x_c = jP.$$

The length of W_j is $w+2w_{\mathrm{g}}=\ell_{\mathrm{c}}$, throughout which $v=-\lambda_{\mathrm{h}}$. The right endpoint of the last W_j is $S-1/2$, so all these intervals stay within the nonconstant part of H. Each W_j extends I_j by w_{g} on both sides. Later we will use this margin to keep every local coordinate in a region with the same amplitude derivative.

Mean objective and client losses. Define

$$a_g(x) = \begin{cases} \sqrt{g}, & x \le 1, \\ \sqrt{g/x}, & 1 \le x \le x_c, \\ \mu_{\mathrm{h}} g, & x_c \le x \le R_f, \\ \mu_{\mathrm{h}} g - (x-R_f)/4, & R_f \le x \le R_f + 4\mu_{\mathrm{h}} g, \\ 0, & x \ge R_f + 4\mu_{\mathrm{h}} g, \end{cases} \qquad f_g(x) = -\int_1^x a_g(y)\,dy.$$

The values at the joins agree; for example, $\sqrt{g/x_c} = g/\sqrt{M_{\mathrm{h}}} = \mu_{\mathrm{h}} g$. On $[1,x_c]$, the derivative has magnitude $\sqrt{g}/(2x^{3/2}) \le \sqrt{g}/2 < 1/4$; the final decreasing segment has slope $-1/4$. Thus a_g is continuous, nonnegative, nonincreasing, and $1/4$-Lipschitz. Since $f_g' = -a_g$ and $a_g = 0$ beyond $R_f + 4\mu_{\mathrm{h}} g$, f_g attains its minimum there. Write $m_g := f_g(R_f + 4\mu_{\mathrm{h}} g)$. The three nonzero integrals on $[1, R_f + 4\mu_{\mathrm{h}} g]$ are

$$\int_1^{x_c} \sqrt{g/x}\,dx = 2\sqrt{M_{\mathrm{h}}} - 2\sqrt{g},$$
$$\int_{x_c}^{R_f} \mu_{\mathrm{h}} g\,dx = \mu_{\mathrm{h}} g(K_{\mathrm{c}}P + 2),$$
$$\int_{R_f}^{R_f+4\mu_{\mathrm{h}} g} a_g(x)\,dx = 2\mu_{\mathrm{h}}^2 g^2.$$

Adding them gives

$$\Delta_g := -m_g = 2\sqrt{M_{\mathrm{h}}} - 2\sqrt{g} + \mu_{\mathrm{h}} g(K_{\mathrm{c}}P + 2) + 2\mu_{\mathrm{h}}^2 g^2, \qquad \Delta_g + 2\sqrt{g} < 2057K.$$

For the last inequality, use $K_{\mathrm{c}} g \le 1$, $P \le 54880K^2 + 2$, $\mu_{\mathrm{h}} = 1/(1000K)$, and $g \le 1$:

$$\Delta_g + 2\sqrt{g} \le 2000K + 54.88K + \frac{4}{1000K} + \frac{2}{10^6 K^2} < 2057K.$$

Define

$$\chi(x) = \begin{cases} 0, & x \le -1, \\ 3(x+1)^2 - 2(x+1)^3, & -1 \le x \le 0, \\ 1, & x \ge 0. \end{cases}$$

Using the notation F_0, Q from Section 4.2, set

$$F_0(X) := \kappa_K \left[m_g + \chi(x(X))(f_g(x(X)) - m_g) + \nu\phi_{\mathrm{H}}(y(X)) \right],$$
$$Q(X) := \kappa_K H(x(X))\psi(y(X)),$$
$$F_\sigma(X) := F_0(X) + \sigma Q(X).$$

For $x(X) \geq 0$, F_σ and its gradient agree with those of $\kappa_K[f_g(x(X)) + \nu\phi_{\mathrm{H}}(y(X)) + \sigma H(x(X))\psi(y(X))]$. This includes $x(X) = 0$, since $\chi'(0) = 0$.

Assumptions and initialization. We first verify global smoothness, including the intervals outside the eventual trajectory. On $-1 < x < 0$, $f_g(x) - m_g = \Delta_g + \sqrt{g}(1-x) < 2057K$, $f_g'(x) = -\sqrt{g}$, and $f_g''(x) = 0$. Differentiating the scalar mean term twice gives

$$\frac{d^2}{dx^2}\{m_g + \chi(x)(f_g(x) - m_g)\} = \chi''(x)(f_g(x) - m_g) - 2\sqrt{g}\,\chi'(x).$$

The bounds $|\chi'| \leq 3/2$ and $|\chi''| \leq 6$ therefore give a scaled derivative-Lipschitz bound $\kappa_K(6 \cdot 2057K + 3\sqrt{g}) < 0.562$ on this transition interval. For $x < -1$ the mean term is constant; for $x > 0$ its scaled derivative is $-\kappa_K a_g(x)$ and is $\kappa_K/4$-Lipschitz. Values and first derivatives agree at the joins, so 0.562 is a global bound for this scalar mean term.

The Hessian entries of $H(x)\psi(y)$, wherever they exist, satisfy

$$|H''(x)\psi(y)| \leq 3\lambda_{\mathrm{h}}, \quad |H'(x)\psi'(y)| \leq \lambda_{\mathrm{h}}, \quad |H(x)\psi''(y)| \leq 2h.$$

The operator norm is bounded by the sum of the absolute entries, which is at most $5\lambda_{\mathrm{h}} + 2h$. The common transverse penalty has gradient-Lipschitz constant ν. Because E_x, E_y are orthonormal, the map $X \mapsto (x(X), y(X))$ has operator norm one. These scalar-coordinate bounds consequently imply

$$\mathrm{Lip}(\nabla F_\sigma) \leq 0.562 + \kappa_K(\nu + 5\lambda_{\mathrm{h}} + 2h) < 0.563 < 1.$$

All first derivatives are continuous across the piecewise joins. Integrating the bounded derivatives on each part of a line segment therefore proves the Lipschitz bound globally, not only where a Hessian is defined or where the algorithm evaluates a gradient.

The global objective satisfies $F \geq F_\star = \kappa_K m_g$, since $\chi \geq 0$, $f_g - m_g \geq 0$, and $\phi_{\mathrm{H}} \geq 0$. Equality is attained at $X_\star = -E_x$. With $B^0 = 0$, $A^0 = e_1^\top$, and the selected minimizing factors $B_\star = -e_1$, $A_\star = e_1^\top$,

$$\|B^0\|_F^2 + \|A^0\|_F^2 = 1, \qquad \|X_\star\|_F = 1, \qquad \|B^0 - B_\star\|_F^2 + \|A^0 - A_\star\|_F^2 = 1.$$

Also

$$\Delta_0 = J(V^0) - F_\star = \kappa_K(\Delta_g + \sqrt{g}) < \frac{2057}{22000} < \frac{1}{10}.$$

Finally, $\nabla F_\sigma(X) - \nabla F(X) = \sigma \nabla Q(X)$ and

$$\nabla Q(X) = \kappa_K[H'(x(X))\psi(y(X))E_x + H(x(X))\psi'(y(X))E_y].$$

Orthonormality of E_x, E_y turns its squared norm into the sum of the squares of these two coefficients. Thus, for every $X \in \mathbb{R}^{2\times 2}$,

$$\frac{1}{2}\sum_{\sigma=\pm 1} \|\nabla F_\sigma(X) - \nabla F(X)\|_F^2 \leq \kappa_K^2[(3\lambda_{\mathrm{h}}/2)^2 + (2h)^2] < (\kappa_K/4)^2 \leq \frac{1}{40000^2}.$$

Thus all stated regularity and initialization bounds are uniform in K and ϵ.

C.2 Exact Local and Server Dynamics

The client-specific terms have opposite signs, so the e_2 components of the local factors cancel at aggregation. Their effect on the e_1 components does not cancel.

Fix any stepsize sequence with $\eta_t \in (0, 1]$. In round t, set $\xi = \kappa_K \eta_t$ and omit the round index from the local notation. Suppose the server factors are $B = be_1$, $A = ae_1^\top$, with $0 \leq b < a$, and write

$$x = ab, \qquad R = a^2 + b^2, \qquad \delta = a^2 - b^2 > 0, \qquad R^2 = 4x^2 + \delta^2, \qquad \rho := \xi R.$$

At round t, these quantities are x_t, R_t, and δ_t; in particular, $R_t = \|V^t\|_F^2$ as defined in Appendix A. Reflection symmetry gives local factors of the form

$$B_{\sigma,e} = b_e e_1 - \sigma z_e e_2, \qquad A_{\sigma,e} = a_e e_1^\top - \sigma w_e e_2^\top.$$

Here $a_0 = a$, $b_0 = b$, and $z_0 = w_0 = 0$. At a local iterate with $x_e = a_e b_e \geq 0$, define

$$d_e = \frac{a_e z_e + b_e w_e}{\sqrt{2}}, \qquad \alpha_e = a_g(x_e) + H'(x_e)\psi(d_e), \qquad T_e = H(x_e)\psi'(d_e) - \nu\phi'_{\mathrm{H}}(d_e). \tag{36}$$

Multiplying the two local factors gives $y(B_{\sigma,e}A_{\sigma,e}) = -\sigma d_e$. The function ψ is odd, its derivative is even, and ϕ'_{H} is odd. Therefore, at this local point,

$$\begin{aligned}\partial_x(F_\sigma/\kappa_K) &= -a_g(x_e) + \sigma H'(x_e)\psi(-\sigma d_e) = -\alpha_e,\\ \partial_y(F_\sigma/\kappa_K) &= \nu\phi'_{\mathrm{H}}(-\sigma d_e) + \sigma H(x_e)\psi'(-\sigma d_e) = \sigma T_e.\end{aligned}$$

Equivalently,

$$\nabla F_\sigma(B_{\sigma,e}A_{\sigma,e}) = \kappa_K \begin{pmatrix} -\alpha_e & \sigma T_e/\sqrt{2} \\ \sigma T_e/\sqrt{2} & 0 \end{pmatrix}.$$

Multiply this matrix by $A_{\sigma,e}^\top$ and multiply it on the left by $B_{\sigma,e}^\top$, as in the factor-gradient formulas of Section 3. The simultaneous updates are

$$b_{e+1} = b_e + \xi\alpha_e a_e + \xi T_e w_e/\sqrt{2}, \qquad z_{e+1} = z_e + \xi T_e a_e/\sqrt{2}, \tag{37}$$

$$a_{e+1} = a_e + \xi\alpha_e b_e + \xi T_e z_e/\sqrt{2}, \qquad w_{e+1} = w_e + \xi T_e b_e/\sqrt{2}. \tag{38}$$

The coefficients in these four recurrences do not depend on σ. Starting from $z_0 = w_0 = 0$, the recurrences therefore also prove the stated reflection form by induction on the local step. Adding the first-coordinate recurrences, and then subtracting them, gives

$$\begin{aligned}a_{e+1} + b_{e+1} &= (1 + \xi\alpha_e)(a_e + b_e) + (\xi T_e/\sqrt{2})(z_e + w_e),\\ a_{e+1} - b_{e+1} &= (1 - \xi\alpha_e)(a_e - b_e) + (\xi T_e/\sqrt{2})(z_e - w_e).\end{aligned}$$

The corresponding second-coordinate recurrences are $z_{e+1} \pm w_{e+1} = z_e \pm w_e + (\xi T_e/\sqrt{2})(a_e \pm b_e)$. Thus the two pairs evolve by the matrices

$$M_{\pm,e} = \begin{pmatrix} 1 \pm \xi\alpha_e & \xi T_e/\sqrt{2} \\ \xi T_e/\sqrt{2} & 1 \end{pmatrix}. \tag{39}$$

After j local steps, define the scalar coefficients

$$L_{\pm,j} = e_1^\top M_{\pm,j-1}\cdots M_{\pm,0}e_1, \quad C_j = \frac{L_{+,j} + L_{-,j}}{2}, \quad D_j = \frac{L_{+,j} - L_{-,j}}{2}.$$

The initial second coordinates of both pairs are zero. Hence $a_j + b_j = L_{+,j}(a + b)$ and $a_j - b_j = L_{-,j}(a - b)$. Adding and subtracting these identities yields

$$b_j = C_j b + D_j a, \qquad a_j = C_j a + D_j b.$$

Multiplying gives

$$x_j = (C_j a + D_j b)(C_j b + D_j a) = (C_j^2 + D_j^2)x + C_j D_j R.$$

The same identities imply

$$a_j^2 + b_j^2 = \tfrac{1}{2}[L_{+,j}^2(a + b)^2 + L_{-,j}^2(a - b)^2].$$

This last identity will bound the retained coordinates after averaging. At communication, the reflected e_2 coordinates cancel, but the e_1 coordinates a_K, b_K agree between the two clients and remain. Expanding $(a^+)^2 - (b^+)^2$ therefore gives the server map

$$b^+ = C_K b + D_K a, \quad a^+ = C_K a + D_K b, \quad \delta^+ = (C_K^2 - D_K^2)\delta.$$

The superscript $+$ denotes the next server iterate. A round is called forward when $x^+ > x$. The coefficients C_j, D_j in this lower-bound proof are scalars; they are not the FedStabLoRA correction states or client displacements.

C.3 Transfer Bounds and Trajectory Invariants

Local matrix products. We bound C_j, D_j and d_j to control the change in the server coordinate. For now, the estimates apply whenever the preceding local iterates satisfy $x_e \geq 0$. We verify this condition by induction below. Equations (32) and (36) give

$$\begin{aligned} |\alpha_e| \leq \alpha_{\max} := \sqrt{g} + \tfrac{3}{2}\lambda_{\mathrm{h}} < \frac{1.02 \cdot 10^{-4}}{K} < 0.02, \\ |T_e| < 0.22, \qquad T_e \geq -\frac{3}{80K}, \qquad T_0 = H(x) \geq h. \end{aligned} \tag{40}$$

For the lower bound on T_e, use $H \leq 2h = 1/5$, $\psi' \geq -1/(16K)$, and $|\phi'_{\mathrm{H}}| \leq 1$ to obtain $T_e \geq -1/(80K) - 1/(40K) = -3/(80K)$. At $e = 0$, $d_0 = 0$, $\psi'(0) = 1$, and $\phi'_{\mathrm{H}}(0) = 0$, so $T_0 = H(x)$.

Lemma C.2 (Elementary matrix bounds)**.** *Every contiguous subproduct of at most K matrices of either sign in* (39) *has its $(1,1)$ entry in $(0.99, 1.01)$. For $1 \leq j \leq K$,*

$$|C_j - 1| \leq K^2\xi^2, \qquad |C_j^2 + D_j^2 - 1| \leq 3K^2\xi^2, \tag{41}$$

$$D_j = \xi \sum_{i=0}^{j-1} \omega_{j,i}\alpha_i, \qquad 0.98 < \omega_{j,i} < 1.03. \tag{42}$$

In particular $L_{\pm,j} > 0$, $C_j > 0$, and $C_j^2 - D_j^2 > 0$.

Proof. **Step 1: bound every subproduct near the identity.** By (40),

$$\|M_{\pm,e} - I\|_{\mathrm{op}} \leq \xi(|\alpha_e| + |T_e|/\sqrt{2}) < \xi(0.02 + 0.22/\sqrt{2}) < 0.18\xi.$$

For any product of $k \leq K$ such matrices, expand each factor as $I + (M_{\pm,e} - I)$ and bound the nonempty products of increments. Its difference from I has norm at most

$$(1 + 0.18\xi)^k - 1 \leq \exp(0.18K\xi) - 1 \leq \exp(0.18/22000) - 1 < 0.01.$$

Every $(1,1)$ entry therefore lies in $(0.99, 1.01)$, including empty products, whose value is one.

Step 2: expand the difference of the two products. Replace the minus matrices by the plus matrices one at a time. This finite telescoping identity gives

$$\begin{aligned} &M_{+,j-1} \cdots M_{+,0} - M_{-,j-1} \cdots M_{-,0} \\ &\quad = \sum_{i=0}^{j-1} M_{+,j-1} \cdots M_{+,i+1}(M_{+,i} - M_{-,i})M_{-,i-1} \cdots M_{-,0}. \end{aligned}$$

Since $M_{+,i} - M_{-,i} = 2\xi\alpha_i e_1 e_1^\top$, taking the $(1,1)$ entry and dividing by two gives (42) with

$$\omega_{j,i} = \big(e_1^\top M_{+,j-1} \cdots M_{+,i+1} e_1\big)\big(e_1^\top M_{-,i-1} \cdots M_{-,0} e_1\big).$$

Step 1 gives $0.99^2 < \omega_{j,i} < 1.01^2$, hence the stated bounds $0.98 < \omega_{j,i} < 1.03$.

Step 3: bound the sum of the two products. Keep the actual coefficients T_i fixed and set all diagonal coefficients α_i to zero. Let C_j^0 denote the resulting $(1,1)$ entry. Expand the original product according to which factors contribute a diagonal term $\pm\xi\alpha_i e_1 e_1^\top$. Terms with an odd number of such terms cancel in the average of the plus and minus products. For

$$a_{\mathrm{ins}} := \xi \sum_{i<j} |\alpha_i|, \qquad v_{\mathrm{ins}} := \xi \sum_{i<j} |T_i|/\sqrt{2},$$

the sum of the norms of terms with r diagonal factors is at most $e^{v_{\mathrm{ins}}} a_{\mathrm{ins}}^r / r!$. This follows by bounding the remaining off-diagonal factors by their product of norms and by expanding $(\sum_{i<j} \xi|\alpha_i|)^r$. Summing the terms with even $r \geq 2$ gives

$$\begin{aligned} |C_j - C_j^0| &\leq e^{v_{\mathrm{ins}}}(\cosh a_{\mathrm{ins}} - 1) \\ &\leq a_{\mathrm{ins}}^2 e^{a_{\mathrm{ins}} + v_{\mathrm{ins}}} \leq 2a_{\mathrm{ins}}^2. \end{aligned} \tag{43}$$

For the final inequality, use $a_{\text{ins}} + v_{\text{ins}} \le (0.02 + 0.22/\sqrt{2})K\xi < 1/100$.

When the diagonal terms are zero, every matrix has the form $I + (\xi T_i/\sqrt{2})\left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$. A product contributes to its $(1,1)$ entry only through an even number of off-diagonal factors. The same expansion therefore gives $|C_j^0 - 1| \le \cosh v_{\text{ins}} - 1 \le v_{\text{ins}}^2$, as $v_{\text{ins}} < 1$. Consequently,

$$|C_j - 1| \le 2a_{\text{ins}}^2 + v_{\text{ins}}^2 \le [2(0.02)^2 + (0.22)^2/2]K^2\xi^2 \le K^2\xi^2.$$

Step 2 also gives $|D_j| \le 1.03(0.02)K\xi$. Expanding $C_j^2 - 1 = 2(C_j - 1) + (C_j - 1)^2$ now yields

$$\begin{aligned} |C_j^2 + D_j^2 - 1| &\le 2K^2\xi^2 + K^4\xi^4 + [1.03(0.02)]^2K^2\xi^2 \\ &< 3K^2\xi^2, \end{aligned}$$

using $K^2\xi^2 \le 22000^{-2}$. Finally, $C_j = (L_{+,j} + L_{-,j})/2 > 0$ and $C_j^2 - D_j^2 = L_{+,j}L_{-,j} > 0$ by Step 1. □

Lemma C.3 (Bounds on the transverse coordinate)**.** *For $\rho = \xi R$ and $1 \le j \le K$,*

$$\frac{\rho}{40} \le d_j \le K\rho.$$

In particular $\rho \ge 80$ implies $d_j \ge 2$, and $\rho \ge U = 40R_y$ implies $d_j \ge R_y$, for each noninitial local iterate.

Proof. By Lemma C.2,

$$0.99(a \pm b) < a_i \pm b_i < 1.01(a \pm b) \qquad (i \ge 1).$$

Both initial quantities $a + b$ and $a - b$ are positive. The second-coordinate recursion gives the exact identity

$$z_j \pm w_j = \frac{\xi}{\sqrt{2}} \sum_{i=0}^{j-1} T_i(a_i \pm b_i).$$

The initial term is at least $0.1(a \pm b)$. Each later negative term is bounded below by $-1.01[3/(80K)](a \pm b)$; positive terms can only increase this lower bound. Hence

$$\begin{aligned} z_j \pm w_j &\ge \frac{\xi}{\sqrt{2}} \left[0.1 - 1.01\frac{3(j-1)}{80K}\right](a \pm b) \\ &\ge \frac{0.062125\xi}{\sqrt{2}}(a \pm b) > 0. \end{aligned}$$

For the upper bound, $|T_i| < 0.22$ and the first-coordinate bounds give $|z_j \pm w_j| \le 0.2222K\xi(a \pm b)/\sqrt{2}$. Substitute both bounds into

$$d_j = \frac{(a_j + b_j)(z_j + w_j) + (a_j - b_j)(z_j - w_j)}{2\sqrt{2}}.$$

Since $(a + b)^2 + (a - b)^2 = 2R$, this gives

$$d_j \ge \frac{0.99 \cdot 0.062125}{2}\xi R = 0.030751875\rho > \rho/40,$$

and

$$d_j \le \frac{1.01 \cdot 0.2222}{2}K\xi R = 0.112211K\rho < K\rho.$$

The two threshold implications follow by substitution of $\rho \ge 80$ and $\rho \ge 40R_y$ into the lower bound. □

Expansion and factor imbalance.

Lemma C.4 (Nonexpansion when $\rho \ge 80$)**.** *If $\rho \ge 80$, then $C_j \le 1$ for $1 \le j \le K$. Consequently $(\delta^+)^2 \le \delta^2$. If $\rho \ge U$, then additionally*

$$C_K^2 + D_K^2 - 1 \le -10^{-3}\xi^2, \qquad |C_j - 1| \le 0.01\xi^2 \quad (1 \le j \le K). \tag{44}$$

Proof. **Step 1: determine the signs of the transverse coefficients.** If $\rho \geq 80$, Lemma C.3 gives $d_i \geq 2$ for all noninitial local iterates. Thus $\psi'(d_i) \leq 0$ and $\phi'_{\mathrm{H}}(d_i) = 1$. Combining this with (40) gives

$$-\frac{3}{80K} \leq T_i \leq -\nu = -\frac{1}{40K} \qquad (i \geq 1).$$

Define $S_j = \sum_{i=1}^{j-1}(-T_i)$, with $S_1 = 0$, and $t_0 = \xi H(x)/\sqrt{2}$. For $j \geq 2$, $(j-1)/(40K) \leq S_j \leq 3/80$.

Step 2: bound the product with its diagonal terms removed. Set $v_i = -\xi T_i/\sqrt{2} \geq 0$ for $i \geq 1$. The zero-diagonal matrices commute and have common eigenvectors $(1,1)$ and $(1,-1)$. Therefore

$$C_j^0 = \frac{(1+t_0)\prod_{i=1}^{j-1}(1-v_i) + (1-t_0)\prod_{i=1}^{j-1}(1+v_i)}{2}.$$

Expanding the products gives one plus the even elementary symmetric terms in the v_i, minus t_0 times the odd terms. The odd sum is at least $\sum_i v_i = \xi S_j/\sqrt{2}$. The even sum beyond one is at most

$$\sum_{r\geq 2}\frac{(\sum_i v_i)^r}{r!} \leq \frac{1}{2}(\sum_i v_i)^2 e^{\sum_i v_i} = \frac{\xi^2 S_j^2}{4} e^{\xi S_j/\sqrt{2}}.$$

Since the exponential is below 1.01 and $H(x) \geq 0.1$,

$$\begin{aligned} C_j^0 &\leq 1 - \xi^2 S_j\left(\frac{H(x)}{2} - \frac{1.01 S_j}{4}\right) \\ &\leq 1 - \xi^2 S_j\left(0.05 - \frac{1.01(3/80)}{4}\right). \end{aligned}$$

The bracket is 0.04053125. Using the lower bound on S_j gives

$$C_j^0 \leq 1 - 0.00101328125\frac{j-1}{K}\xi^2. \tag{45}$$

Step 3: restore the diagonal terms. By (43) and (40),

$$|C_j - C_j^0| \leq 2j^2\xi^2\alpha_{\max}^2 \leq 2(1.02\cdot 10^{-4})^2(j/K)^2\xi^2.$$

For $2 \leq j \leq K$, we have $j^2 \leq 2K(j-1)$. Thus the coefficient of ξ^2 in this error is at most $4(1.02\cdot 10^{-4})^2(j-1)/K$, strictly smaller than the subtracted coefficient in (45). It follows that $C_j \leq 1$ for $j \geq 2$; directly $C_1 = 1$. Since $0 < C_K^2 - D_K^2 \leq C_K^2 \leq 1$, $\delta^+ = (C_K^2 - D_K^2)\delta$ implies $(\delta^+)^2 \leq \delta^2$.

Step 4: strengthen the bound when $\rho \geq U$. Now Lemma C.3 gives $d_i \geq R_y$ for $i \geq 1$. Hence $T_i = -\nu$ and $S_K = (K-1)/(40K) \in [1/80, 1/40]$. The calculation in Step 2 becomes

$$\begin{aligned} C_K^0 &\leq 1 - \xi^2\frac{1}{80}\left(0.05 - \frac{1.01}{160}\right) \\ &= 1 - 0.00054609375\xi^2. \end{aligned}$$

The restored diagonal error is less than $2.081\cdot 10^{-8}\xi^2$, so $C_K \leq 1 - 0.00054\xi^2$. Also $D_K^2 \leq (1.03K\alpha_{\max})^2\xi^2 < 1.104\cdot 10^{-8}\xi^2$. Consequently,

$$\begin{aligned} C_K^2 + D_K^2 - 1 &\leq [-0.00108 + (0.00054)^2\xi^2 + 1.104\cdot 10^{-8}]\xi^2 \\ &\leq -10^{-3}\xi^2, \end{aligned}$$

where $\xi \leq 1$ is sufficient for the last bound. For any prefix, Step 2 and $C_j^0 \leq 1$ imply

$$0 \leq 1 - C_j^0 \leq t_0(\sum_i v_i)e^{\sum_i v_i} \leq \tfrac{1}{2}(1.01)H(x)S_j\xi^2 \leq 0.002525\xi^2.$$

Adding the same diagonal error proves $|C_j - 1| \leq 0.01\xi^2$. For $j = 1$, this holds directly since $C_1 = C_1^0 = 1$. □

Lemma C.5 (One-round bound on factor imbalance)**.** *In any round with nonnegative server factors and $x_e \geq 0$ at every local gradient evaluation,*

$$(\delta^+)^2 \leq \delta^2 + \mathcal{B}_K, \qquad \mathcal{B}_K = 51200K^2.$$

Proof. If $\rho \geq 80$, Lemma C.4 proves the stronger bound $(\delta^+)^2 \leq \delta^2$. Suppose $\rho < 80$. By Lemma C.2, $0 < C_K^2 - D_K^2 \leq C_K^2 \leq (1 + K^2\xi^2)^2$. For $z = K^2\xi^2 \leq 22000^{-2}$, expansion gives $(1+z)^4 - 1 = (4 + 6z + 4z^2 + z^3)z \leq 7z$. Therefore

$$\begin{aligned}(\delta^+)^2 - \delta^2 &= [(C_K^2 - D_K^2)^2 - 1]\delta^2 \\ &\leq 7K^2\xi^2\delta^2 \leq 7K^2\xi^2R^2 = 7K^2\rho^2 \\ &< 7K^2 \cdot 80^2 < 51200K^2.\end{aligned}$$

Here $\delta \leq R$ follows from $R = a^2 + b^2$ and $\delta = a^2 - b^2 > 0$. □

Nonnegative local coordinates and factor bounds. Let

$$\mathcal{H} = 74g^{-3}, \qquad \mathcal{T} = \inf\{t \geq 0 : x_t > Z\}.$$

We prove, by induction on rounds and local steps, that for every integer $t < \min(\mathcal{T}, \mathcal{H})$, and at the endpoint of such a round,

$$B^t = b_t e_1, \quad A^t = a_t e_1^\top, \quad 0 \leq b_t < a_t, \quad R_t \geq 1, \quad 0 < \delta_t^2 \leq 1 + \mathcal{B}_K t.$$

This is a nested induction. Assume the server bounds at round t. At local step j, assume the required coordinate bounds only for $e < j$. The coefficients of the first j update matrices are then defined at nonnegative coordinates, so the preceding lemmas apply to that prefix and bound its endpoint x_j. We next verify that this endpoint obeys the same coordinate bounds. Only after all K local steps have been treated do we update the server bounds. In particular, none of the prefix estimates assumes in advance the nonnegativity of the endpoint it is being used to prove.

At the start of a round, the induction hypothesis gives

$$R \leq 2x + \delta, \qquad \delta \leq \sqrt{1 + 74\mathcal{B}_K g^{-3}}, \qquad 0 \leq x \leq Z. \tag{46}$$

Initially these assertions hold with $(b, a, x, R, \delta) = (0, 1, 0, 1, 1)$.

Local displacement when $\rho < U$. When $\rho < U$, the estimates in Equations (41) and (42) imply

$$\begin{aligned}|x_j - x| &\leq 3K^2\xi^2 x + 0.021K\xi R &(47)\\ &\leq (1.5K^2\kappa_K + 0.021K)U < 0.022KU < J_0. &(48)\end{aligned}$$

To see the second line, use $x \leq R/2$, $\xi \leq \kappa_K$, and $\xi R = \rho < U$ in the first line. The coefficient satisfies $1.5K\kappa_K + 0.021 < 0.022$. The bound for x_j requires only $x_e \geq 0$ for $e < j$. Also $J_0 < x_c/4$: indeed, $J_0 g/M_{\mathrm{h}} \leq 1960/(10^{14}K^2) < 1/4$. This last inequality ensures that the local-displacement estimate cannot cross zero from a start at $x \geq x_c/2$.

Rounds starting at $x < x_c/2$. We show inductively that $0 \leq x_j < x_c$ for $j = 1, \ldots, K$. As long as earlier local iterates have this property, their $\alpha_i = a_g(x_i)$ are positive. Equation (42) gives $D_j > 0$, and $C_j > 0$, so $b_j \geq 0$, $a_j > 0$, and $x_j \geq 0$. If $\rho < U$, (48) and $J_0 < x_c/4$ give $x_j < 3x_c/4$. If $\rho \geq U$, Lemma C.4 gives $C_j \leq 1$. The initial coefficient satisfies $|\alpha_0| \leq \sqrt{g}$; at each subsequent local query, $\psi(d_i) = \Gamma g$ gives $|\alpha_i| \leq \sqrt{g} + \lambda_{\mathrm{h}}\Gamma g \leq 2\sqrt{g}$. Thus $|D_j| \leq 3K\xi\sqrt{g}$. Since $C_j \leq 1$, $x \leq R/2$, and $K\xi\sqrt{g} < 1/100$, the exact server-coordinate formula gives

$$\begin{aligned}x_j - x &\leq D_j^2 x + |D_j| R \\ &\leq \left[\tfrac{9}{2}(K\xi\sqrt{g})^2 + 3K\xi\sqrt{g}\right] R \leq 4K\xi\sqrt{g}\, R.\end{aligned}$$

By (46),

$$g(x_j - x) \leq 4K\kappa_K\left(M_{\mathrm{h}}\sqrt{g} + \sqrt{g^3 + 74\mathcal{B}_K}\right) < M_{\mathrm{h}}/4.$$

For an explicit check, $M_{\mathrm{h}}\sqrt{g} \leq 100K$ and $\sqrt{g^3 + 74\mathcal{B}_K} < 2000K$. Thus the displayed right-hand side is less than $4(2100K)/22000 < K$, whereas $M_{\mathrm{h}}/4 = 250000K^2$. This proves the final

inequality uniformly for all $K \geq 2$. Thus again $x_j < 3x_c/4$. This closes the local induction and proves in particular that a round starting at $x < x_c/2$ ends before x_c.

Rounds starting at $x \geq x_c/2$. If $\rho < U$, (48) gives $x_j > x - J_0 > x_c/4 > 0$, which completes this induction step. If $\rho \geq U$, inductively suppose earlier coordinates are at least $x_c/4$. Their slopes satisfy $a_g(x_i) \leq 2\mu_{\mathrm{h}} g$, and $\lambda_{\mathrm{h}}\Gamma = 10K\mu_{\mathrm{h}} = 0.01$. Hence

$$|\alpha_i| \leq 0.011g, \qquad |D_j| \leq 0.012K\xi g, \qquad 0.99 < C_j \leq 1. \tag{49}$$

Using (46),

$$\frac{gR}{x} \leq 2g + \frac{2\sqrt{g}}{M_{\mathrm{h}}}\sqrt{g^3 + 74\mathcal{B}_K} < 1.$$

For the last inequality in the preceding display, the same bound $\sqrt{g^3 + 74\mathcal{B}_K} < 2000K$ gives $2g + (2\sqrt{g}/M_{\mathrm{h}})\sqrt{g^3 + 74\mathcal{B}_K} \leq 2/(10^8K^2) + 0.4/(10^6K^2) < 1$. Since $C_j > 0.99$, $C_j \leq 1$, and $|D_j| \leq 0.012K\xi g$, the exact expression for x_j therefore gives

$$x_j \geq 0.9801x - 0.012K\xi gR > 0.97x > x_c/4,$$

since $0.012K\kappa_K < 0.01$. This closes the local induction here too. The sum $a_j + b_j$ remains positive; $x_j > 0$ and positivity of both matrix-product multipliers imply $a_j > b_j > 0$.

Factor norm and imbalance after aggregation. Since $C_K^2 - D_K^2 > 0$, the server update preserves $\delta > 0$, and Lemma C.5 bounds $(\delta^+)^2$. If $R > 2$, the matrix-product bounds and the identity for $a_j^2 + b_j^2$ give $R^+ \geq 0.99^2 R > 1$. If $1 \leq R \leq 2$, that identity gives $a_j^2 + b_j^2 \leq 1.01^2 R \leq 2.0402$, and thus $x_j = a_j b_j \leq (a_j^2 + b_j^2)/2 \leq 1.0201 < x_c$. Lemma C.3 also gives $d_j \leq 2K\kappa_K < 1$. Thus $H = h$, $\alpha_j > 0$, and $T_j = h - \nu d_j > 0$ at every local gradient evaluation. The original four-coordinate updates then have nonnegative coefficients and nonnegative factors; a_j, b_j cannot decrease. Therefore $R^+ \geq R \geq 1$ in this case as well. This proves the induction claims for the next server iterate as well.

C.4 Per-Round Progress in Slow Intervals

Bounds on jumps into slow intervals. For a round with $x \geq x_c/2$ and $t < \min\{\mathcal{T}, \mathcal{H}\}$, $x \leq Z < R_f = M_{\mathrm{h}}/g + K_{\mathrm{c}}P + 2 < 2M_{\mathrm{h}}/g$, since $K_{\mathrm{c}} \leq g^{-1}$ and $P + 2g \leq 54880K^2 + 4 < M_{\mathrm{h}}$. Thus

$$\frac{g^2R^2}{x} = 4g^2x + \frac{g^2\delta^2}{x} \leq 8M_{\mathrm{h}}g + \frac{2(g^3 + 74\mathcal{B}_K)}{M_{\mathrm{h}}} < 8. \tag{50}$$

Indeed, $M_{\mathrm{h}}g \leq 1/100$, $g \leq 1$, and $M_{\mathrm{h}} \geq 4 \cdot 10^6$ give

$$8M_{\mathrm{h}}g + \frac{2(g^3 + 74\mathcal{B}_K)}{M_{\mathrm{h}}} \leq 0.08 + \frac{2}{M_{\mathrm{h}}} + \frac{2 \cdot 74 \cdot 51200}{10^6} < 8.$$

If $\rho \geq U$, the bounds in (49) imply

$$x^+ - x \leq -10^{-3}\xi^2 x + 0.012K\xi gR. \tag{51}$$

Here the negative term uses (44), not just the bounds on D_K. Since $x > 0$, complete the square in ξ:

$$\begin{aligned} -10^{-3}x\xi^2 + 0.012KgR\xi &= -10^{-3}x\left(\xi - \frac{0.012KgR}{2 \cdot 10^{-3}x}\right)^2 + \frac{(0.012KgR)^2}{4 \cdot 10^{-3}x} \\ &\leq 0.036K^2\frac{g^2R^2}{x} < 0.288K^2 < J_0. \end{aligned}$$

The final strict inequality uses (50). Together with (48) and the case $x < x_c/2$,

$$x < x_c/2 \implies x^+ < x_c, \qquad x_c/2 \leq x \leq Z \implies x^+ - x < J_0. \tag{52}$$

Suppose now a round starts in I_j and is forward. For $\rho < U$, all its local coordinates are within $J_0 < w_{\mathrm{g}}$ of its starting coordinate. For $\rho \geq U$, forward movement in (51) forces

$$\xi x < 12KgR. \tag{53}$$

For the entire local prefix, (44) gives $|C_i - 1| \le 0.01\xi^2$. Moreover, $0.012Kg \le 0.012/(10^8 K) < 10^{-4}$. Thus expansion of $C_i^2 - 1$ and the bound $|D_i| \le 0.012K\xi g$ imply $|C_i^2 + D_i^2 - 1| \le 0.03\xi^2$. Substitution in the exact expression for x_i gives

$$|C_i^2 + D_i^2 - 1| \le 0.03\xi^2, \qquad |x_i - x| \le 0.03\xi^2 x + 0.012K\xi g R.$$

Using (53) and (50),

$$|x_i - x| \le (0.03 \cdot 12^2 + 0.012 \cdot 12)K^2 \frac{g^2 R^2}{x} < 35.712K^2 < w_{\mathrm{g}}.$$

Thus, whenever a round starts in I_j and has $x^+ > x$, every local coordinate x_i lies in W_j. Rounds with $x^+ \le x$ need no such bound.

Progress within a slow interval. Consider a round that starts in I_j and increases x. All its local coordinates lie in W_j, where $H' = -\lambda_{\mathrm{h}}$. A large value of $\psi(d_i)$ therefore makes the coefficient α_i negative. We show that the round can increase x only if ξ is small. Since all local coordinates lie in W_j,

$$\alpha_0 = \mu_{\mathrm{h}} g, \qquad \alpha_i = \mu_{\mathrm{h}} g - \lambda_{\mathrm{h}} \psi(d_i) \quad (i \ge 1).$$

If $\rho \ge 40\Gamma g$, then $d_i \ge \Gamma g$ for every $i \ge 1$. Consequently $\alpha_i \le (1 - 10K)\mu_{\mathrm{h}} g$ for $i \ge 1$. In (42), the initial coefficient is positive and the other $K - 1$ coefficients are negative. Upper-bound the first weight by 1.03 and lower-bound the weights multiplying negative terms by 0.98. This gives

$$D_K \le \xi\mu_{\mathrm{h}} g[1.03 - 0.98(10K - 1)(K - 1)] \tag{54}$$

$$\le -3K^2 \mu_{\mathrm{h}} \xi g < 0. \tag{55}$$

The last inequality holds for every $K \ge 2$: after expansion, its bracket is $-9.8K^2 + 10.78K + 0.05 \le -3K^2$. For $\rho < U$, the starting point lies in $I_j \subset (x_c, R_f)$, so $R \ge 2x \ge 2M_{\mathrm{h}}/g$. Therefore

$$\xi \le \frac{Ug}{2M_{\mathrm{h}}}, \qquad C_K - 1 \le K^2\xi^2 \le \frac{U}{2\sqrt{M_{\mathrm{h}}}} K^2 \mu_{\mathrm{h}} \xi g \le 0.98K^2 \mu_{\mathrm{h}} \xi g.$$

The last bound uses $U/(2\sqrt{M_{\mathrm{h}}}) \le 0.98$. Combining it with $D_K \le -3K^2\mu_{\mathrm{h}}\xi g$ gives $C_K + D_K < 1$. For $\rho \ge U$, the bounds $C_K \le 1$ and $D_K < 0$ give the same conclusion. In both cases, $C_K + D_K = L_{+,K} > 0$ by Lemma C.2. Thus $0 < C_K + D_K < 1$. Since $C_K > 0$, $D_K < 0$, and $R \ge 2x$,

$$x^+ = (C_K^2 + D_K^2)x + C_K D_K R \le (C_K + D_K)^2 x < x,$$

a contradiction. Every forward round starting in I_j therefore satisfies

$$\rho < 40\Gamma g, \qquad \xi < 20\Gamma g^2/M_{\mathrm{h}}. \tag{56}$$

These bounds follow from $x^+ > x$; they are not additional restrictions on the allowed stepsizes $\eta_t \in (0, 1]$.

Since $d_i \ge 0$, all local coefficients in this round satisfy $\alpha_i \le \mu_{\mathrm{h}} g$. Thus $D_K \le 1.03K\mu_{\mathrm{h}}\xi g$ and $C_K D_K R \le 1.05K\mu_{\mathrm{h}}\xi g R$ (a negative D_K only helps). Use $x \le Z < 2M_{\mathrm{h}}/g$ and the two bounds $\xi < 20\Gamma g^2/M_{\mathrm{h}}$ and $\xi R = \rho < 40\Gamma g$. They bound the two progress terms separately:

$$3K^2\xi^2 x \le \frac{2400K^2\Gamma^2}{M_{\mathrm{h}}} g^3, \qquad 1.05K\mu_{\mathrm{h}}\xi g R \le 42K\mu_{\mathrm{h}}\Gamma g^2.$$

Consequently,

$$x^+ - x \le 3K^2\xi^2 x + 1.05K\mu_{\mathrm{h}}\xi g R \tag{57}$$

$$\le 2400K^2\Gamma^2 {M_{\mathrm{h}}}^{-1} g^3 + 42K\mu_{\mathrm{h}}\Gamma g^2 \tag{58}$$

$$\le 43K\mu_{\mathrm{h}}\Gamma g^2. \tag{59}$$

For the last step, use $2400K\Gamma g/\sqrt{M_{\mathrm{h}}} = 2.4\Gamma g < 1$. Nonforward rounds satisfy the same upper bound automatically. Finally

$$C_{\mathrm{prog}} := 43K\mu_{\mathrm{h}}\Gamma = 0.1204J_0 + 0.0043 < 0.121J_0.$$

C.5 Communication-Round Lower Bound

We use (52) and (59) to count progress without assuming that x_t is monotone. From each slow interval $I_j = [\ell_j, \ell_j + w]$, remove the leftmost segment of length J_0 and define

$$K_j = [\ell_j + J_0, \ell_j + w].$$

Let $\Phi(x)$ be the total length of these remaining intervals to the left of x:

$$\Phi(x) := \int_{-\infty}^{x} \mathbf{1}_{\cup_{j=1}^{K_c} K_j}(u)\, du.$$

It is nondecreasing and 1-Lipschitz.

Lemma C.6 (One-round increase of Φ)**.** *For every round $t < \min\{\mathcal{T}, \mathcal{H}\}$,*

$$\Phi(x_{t+1}) - \Phi(x_t) \le C_{\mathrm{prog}} g^2.$$

Proof. If $x_{t+1} \le x_t$, monotonicity of Φ makes the left-hand side nonpositive. Suppose therefore that the round is forward.

If $x_t \in I_j$ for some j, the 1-Lipschitz property and (59) give

$$\Phi(x_{t+1}) - \Phi(x_t) \le x_{t+1} - x_t \le C_{\mathrm{prog}} g^2.$$

It remains to consider a forward round starting outside all slow intervals. If $x_t < x_c/2$, (52) gives $x_{t+1} < x_c$. Every K_j lies strictly to the right of x_c, so the potential increment is zero. If $x_t \ge x_c/2$, then $x_{t+1} < x_t + J_0$. For any slow interval lying to the right of x_t, being outside I_j implies $x_t < \ell_j$, and hence

$$x_{t+1} < x_t + J_0 < \ell_j + J_0 = \inf K_j.$$

All slow intervals lying to the left have already ended before x_t. Thus this round cannot enter any new K_j and again has zero potential increment. These cases exhaust all possibilities. □

Lemma C.7 (Lower bound on the exit time)**.** *With $\mathcal{T} = \inf\{t \ge 0 : x_t > Z\}$ and $\mathcal{H} = 74g^{-3}$ as in Section C.3, one has $\mathcal{T} \ge \mathcal{H}$.*

Proof. Assume for contradiction that $\mathcal{T} < \mathcal{H}$. Since $x_0 = 0$ lies to the left of all the intervals K_j, $\Phi(x_0) = 0$. At the first exit, $x_{\mathcal{T}} > Z = \ell_{K_c} + w$, so every interval K_j lies to the left of $x_{\mathcal{T}}$ and

$$\Phi(x_{\mathcal{T}}) = \sum_{j=1}^{K_c} |K_j| = K_c(w - J_0) = 9 J_0 K_c.$$

Telescoping Lemma C.6 over $t = 0, \dots, \mathcal{T} - 1$ yields

$$\mathcal{T}\, C_{\mathrm{prog}} g^2 \ge 9 J_0 K_c.$$

The intervals K_j are disjoint because consecutive slow intervals start $P > w$ apart. Their lengths therefore add as used above. Also, the floor inequality $\lfloor u \rfloor \ge u - 1$ and $g \le g_0$ give

$$K_c = \lfloor g^{-1} \rfloor \ge g^{-1} - 1 = \frac{1-g}{g} \ge \frac{1-g_0}{g}.$$

Using $C_{\mathrm{prog}} < 0.121 J_0$, $M_h = 10^6 K^2$, $K \ge 2$, and $g_0 = (100 M_h)^{-1}$, we obtain

$$\mathcal{T} > (1 - g_0) \frac{9}{0.121} g^{-3} \tag{60}$$

$$> 74 g^{-3} = \mathcal{H}, \tag{61}$$

The final constant comparison follows from $g_0 = 1/(10^8 K^2) \le 1/(4 \cdot 10^8)$ and $9/0.121 > 74.38$. This contradicts $\mathcal{T} < \mathcal{H}$. All inequalities used in the sum concern rounds strictly before $\mathcal{T}$, where the trajectory bounds have already been verified; no bound after exit is needed. □

Factor-gradient lower bound. By Lemma C.7, the bounds in Section C.3 hold for every integer $t < \mathcal{H}$. At each such communication round,

$$x_t \geq 0, \qquad y(B^t A^t) = 0, \qquad R_t \geq 1, \qquad x_t \leq Z < R_f.$$

Since $x_t \geq 0$, $\chi(x_t) = 1$ and $\chi'(x_t) = 0$. The client-specific terms cancel in F, and $\phi'_{\mathrm{H}}(0) = 0$. It follows that $\nabla F(B^t A^t) = -\kappa_K a_g(x_t) E_x$. As $B^t = b_t e_1$ and $A^t = a_t e_1^\top$, the two factor gradients are $-\kappa_K a_g(x_t) a_t e_1$ and $-\kappa_K a_g(x_t) b_t e_1^\top$, respectively. Their squared norms add to $\kappa_K^2 a_g(x_t)^2 (a_t^2 + b_t^2)$, so

$$\|\nabla J(V^t)\|_F = \kappa_K a_g(x_t) \sqrt{R_t}.$$

The lower bound now follows in each of the three coordinate ranges:

$$a_g(x_t)\sqrt{R_t} \geq \begin{cases} \sqrt{g}, & 0 \leq x_t \leq 1 \quad (R_t \geq 1), \\ \sqrt{g/x_t}\sqrt{2x_t} = \sqrt{2g}, & 1 \leq x_t \leq M_{\mathrm{h}}/g, \\ \mu_{\mathrm{h}} g \sqrt{2M_{\mathrm{h}}/g} = \sqrt{2g}, & M_{\mathrm{h}}/g \leq x_t \leq Z. \end{cases}$$

The last two cases use $R_t \geq 2x_t$, and the last also uses $\mu_{\mathrm{h}} = {M_{\mathrm{h}}}^{-1/2}$. Thus in all three regions,

$$\|\nabla J(V^t)\|_F \geq \kappa_K \sqrt{g} = \frac{101}{100}\epsilon > \epsilon, \qquad t < \mathcal{H}.$$

It follows that

$$\|\nabla J(V^t)\|_F > \epsilon \quad \text{for } 0 \leq t < \mathcal{H} = 74 g^{-3} = \frac{74}{(22220K)^6}\epsilon^{-6},$$

which proves Theorem C.1 and Theorem 4.2. The bound holds for every integer round index below this threshold.

The objectives depend on (K, ϵ), not on the stepsize sequence. Since all estimates hold for each sequence in $(0,1]^{\mathbb{N}_0}$, they apply in particular to every deterministic rule that chooses the next stepsize from the optimization history. No monotonicity or positive lower bound on the stepsizes is required.

D Proof of the Full-Participation Guarantee for FedStabLoRA

We first bound the error in $D_i^t \approx -s[\nabla J_i(V^t) - C_i^t]$ for the parameters in (4). Combining this bound with the correction-state update gives a decrease in a weighted sum of the objective gap and the correction error. The same inequality bounds factor movement and verifies the local smoothness conditions at every round.

D.1 Local Update Approximation

We use $\eta > 0$, $\mu \geq 0$, $0 \leq \eta\mu \leq 1$, and $s = \eta \sum_{k=0}^{K-1} (1 - \eta\mu)^k$ as in (4). The following constants are used in this section and Appendix E:

$$\delta_0 = \frac{1}{10}, \qquad \rho_* = \frac{1}{9}, \qquad \kappa = \frac{10}{9}.$$

We state the local lemma for a differentiable function f_i. In the original algorithm $f_i = J_i$; the same lemma will also be used for the auxiliary objectives in Appendix E.

Lemma D.1 (Local iterates and the returned update)**.** *Let $b > 0$ and let ∇f_i be ℓ-Lipschitz on $\{U : \|U - V\|_F \leq b\}$. Fix C_i and put $g_i = \nabla f_i(V) - C_i$. Suppose the parameters satisfy* (4) *and*

$$\delta := s\ell \leq \delta_0, \qquad \kappa s \|g_i\|_F \leq b/2. \tag{62}$$

Starting at $V_{i,0} = V$, take K steps $V_{i,e+1} = V_{i,e} - \eta[\nabla f_i(V_{i,e}) - C_i + \mu(V_{i,e} - V)]$. Then all these iterates lie in the ball, and

$$\max_{0 \leq e \leq K} \|V_{i,e} - V\|_F \leq \frac{s}{1-\delta}\|g_i\|_F \leq \kappa s \|g_i\|_F. \tag{63}$$

For $D_i = V_{i,K} - V$, define $a_i := g_i + D_i/s$. Then

$$D_i = -s g_i + s a_i, \qquad \|a_i\|_F \leq \rho_* \|g_i\|_F. \tag{64}$$

More precisely,

$$\|a_i\|_F \leq \frac{\delta}{1-\delta}\|g_i\|_F.$$

If ∇f_i is globally ℓ-Lipschitz, the radius condition in (62) *can be omitted.*

Proof. **Step 1: expand the local recursion.** Put $w = 1 - \eta\mu \in [0, 1]$ and $D_{i,e} = V_{i,e} - V$, so $D_{i,0} = 0$. Adding and subtracting $\nabla f_i(V)$ in the local update gives

$$D_{i,e+1} = wD_{i,e} - \eta g_i - \eta[\nabla f_i(V + D_{i,e}) - \nabla f_i(V)].$$

Repeated substitution, with an empty sum interpreted as zero, yields

$$\begin{aligned} D_{i,e} &= -s_e g_i - \eta \sum_{k=0}^{e-1} w^{e-1-k}[\nabla f_i(V + D_{i,k}) - \nabla f_i(V)], \\ s_e &:= \eta \sum_{k=0}^{e-1} w^k. \end{aligned} \tag{65}$$

By the definition of s, we have $s_K = s$. Since $w \in [0, 1]$, we also have $0 \le s_e \le s$ for every $0 \le e \le K$. When $\mu = 0$, $w = 1$, $s_e = e\eta$, and $s_K = K\eta = s$. The expansion uses only finite sums, so it does not require $w < 1$.

Step 2: verify that the local iterates stay in the ball. The bound holds at $e = 0$. Suppose it holds for $k < e$, namely $\|D_{i,k}\|_F \le s\|g_i\|_F/(1 - \delta)$. Since $\delta \le \delta_0 = 1/10$, $(1 - \delta)^{-1} \le 10/9 = \kappa$. The radius condition in (62) therefore places these earlier iterates inside the ball where the gradient is ℓ-Lipschitz. Taking norms in (65) gives

$$\begin{aligned} \|D_{i,e}\|_F &\le s_e\|g_i\|_F + \eta \sum_{k=0}^{e-1} w^{e-1-k}\ell\|D_{i,k}\|_F \\ &\le s_e \left(1 + \frac{s\ell}{1 - \delta}\right) \|g_i\|_F = \frac{s_e}{1 - \delta}\|g_i\|_F \\ &\le \frac{s}{1 - \delta}\|g_i\|_F \le \kappa s\|g_i\|_F \le b/2. \end{aligned}$$

This proves the next bound and the neighborhood condition together. The induction proves (63) for every local iterate, including the returned endpoint. With a globally Lipschitz gradient, the same argument does not need a radius condition.

Step 3: compare the endpoint with $-sg_i$. Set $e = K$ in (65) and use $a_i = g_i + D_i/s$ and $s_K = s$:

$$a_i = -\frac{\eta}{s} \sum_{k=0}^{K-1} w^{K-1-k}[\nabla f_i(V + D_{i,k}) - \nabla f_i(V)].$$

Applying Step 2 and $\delta \le 1/10$ yields

$$\|a_i\|_F \le \frac{\delta}{1 - \delta}\|g_i\|_F \le \frac{1}{9}\|g_i\|_F = \rho_*\|g_i\|_F.$$

This includes $\mu = 0$: the preceding induction uses only $w \ge 0$ and $s_e \le s$, while the endpoint estimate uses $s_K = s$. Finally, $D_i = -sg_i + sa_i$ follows from the definition of a_i. □

If every local gradient is replaced by its value at V, the local recursion returns $D_i = -sg_i$ exactly.

Feasible parameters. For a prescribed $s > 0$ and $\mu \ge 0$ satisfying $\mu s \le 1$, choose

$$\eta = \begin{cases} s/K, & \mu = 0, \\ \left[1 - (1 - \mu s)^{1/K}\right]/\mu, & \mu > 0. \end{cases}$$

When $\mu = 0$, this gives $\eta > 0$, $\eta\mu = 0$, and $\eta \sum_{k=0}^{K-1}(1 - \eta\mu)^k = K\eta = s$. When $\mu > 0$, we have $0 < \eta\mu \le 1$, and the geometric-sum identity gives

$$\eta \sum_{k=0}^{K-1}(1 - \eta\mu)^k = \frac{1 - (1 - \eta\mu)^K}{\mu} = \frac{1 - (1 - \mu s)}{\mu} = s.$$

The positive-μ branch includes $\mu s = 1$, for which $\eta\mu = 1$.

D.2 LYAPUNOV DESCENT

Under full participation, let $\bar{D}^t = \sum_i p_i D_i^t$. The correction states and their weighted mean satisfy $Y_i^{t+1} = Y_i^t - D_i^t/s$ and $\bar{Y}^{t+1} = \bar{Y}^t - \bar{D}^t/s$. Subtracting gives

$$C_i^{t+1} = C_i^t - \frac{D_i^t - \bar{D}^t}{s}, \qquad C_i^t = Y_i^t - \bar{Y}^t, \qquad \sum_i p_i C_i^t = 0. \tag{66}$$

Define the corrected client gradients and their squared norms by

$$g_i^t := \nabla J_i(V^t) - C_i^t,$$
$$\mathcal{Q}_t := \sum_i p_i \|g_i^t\|_F^2, \qquad \mathcal{E}_t := \sum_i p_i \|g_i^t - \nabla J(V^t)\|_F^2.$$

Since $\sum_i p_i C_i^t = 0$, we have $\sum_i p_i g_i^t = \nabla J(V^t)$ and $\mathcal{Q}_t = \|\nabla J(V^t)\|_F^2 + \mathcal{E}_t$. Thus $\mathcal{E}_t$ measures the difference between the corrected client gradients and the global gradient, all evaluated at V^t. Define

$$\mathcal{L}_t := \Delta_t + \frac{s}{2}\mathcal{E}_t.$$

We add the bound on the change in Δ_t to $s/2$ times the bound on the change in $\mathcal{E}_t$.

Lemma D.2 (Full-participation descent)**.** *Suppose* (62) *holds for every client at round t with $f_i = J_i$, $\ell = \ell(V^t)$, and $b = 1 + \|V^t\|_F$. Then*

$$\mathcal{L}_{t+1} \le \mathcal{L}_t - \frac{s}{4}\mathcal{Q}_t.$$

Proof. At round t, use the endpoint errors $a_i^t := g_i^t + D_i^t/s$ from Lemma D.1. Write $\bar{a}^t = \sum_i p_i a_i^t$ and $\delta = s\ell(V^t)$.

Step 1: bound the objective change. Since $\sum_i p_i g_i^t = \nabla J(V^t)$, averaging (64) gives $\bar{D}^t = s[-\nabla J(V^t) + \bar{a}^t]$. The local-radius estimate also gives, by Jensen's inequality,

$$\|\bar{D}^t\|_F^2 \le \sum_i p_i \|D_i^t\|_F^2 \le \kappa^2 s^2 \mathcal{Q}_t.$$

The local endpoints lie in $\mathcal{B}(V^t)$, and their weighted average V^{t+1} lies there as well. Apply the descent inequality from Lemma A.3 to obtain

$$\begin{aligned}\Delta_{t+1} - \Delta_t &\le \langle \nabla J(V^t), \bar{D}^t \rangle + \tfrac{1}{2}\ell(V^t)\|\bar{D}^t\|_F^2 \\ &\le -s\|\nabla J(V^t)\|_F^2 + s\langle \nabla J(V^t), \bar{a}^t \rangle + \tfrac{1}{2}\kappa^2 s^2 \ell(V^t)\mathcal{Q}_t.\end{aligned}$$

Dividing by s gives

$$\frac{\Delta_{t+1} - \Delta_t}{s} \le -\|\nabla J(V^t)\|_F^2 + \langle \nabla J(V^t), \bar{a}^t \rangle + \frac{\kappa^2 \delta}{2}\mathcal{Q}_t. \tag{67}$$

Step 2: bound the correction error at the new server point. Define $k_i^t = \nabla J_i(V^{t+1}) - \nabla J_i(V^t)$ and $\bar{k}^t = \sum_i p_i k_i^t$. Substituting the centered-state update and then the endpoint identity yields

$$\begin{aligned}g_i^{t+1} &= \nabla J_i(V^t) + k_i^t - C_i^t + (D_i^t - \bar{D}^t)/s \\ &= g_i^t + k_i^t + (-g_i^t + a_i^t) - \bar{D}^t/s \\ &= a_i^t + k_i^t - \bar{D}^t/s.\end{aligned}$$

Its weighted mean is $\nabla J(V^{t+1})$. Subtracting that mean gives

$$g_i^{t+1} - \nabla J(V^{t+1}) = (a_i^t - \bar{a}^t) + (k_i^t - \bar{k}^t).$$

Both server points lie in $\mathcal{B}(V^t)$, so $\|k_i^t\|_F \le \ell(V^t)\|\bar{D}^t\|_F$. Consequently,

$$\sum_i p_i \|k_i^t\|_F^2 \le \kappa^2 \delta^2 \mathcal{Q}_t, \qquad \sum_i p_i \|a_i^t\|_F^2 \le \rho_*^2 \mathcal{Q}_t.$$

Expand the preceding centered sum of squares and apply weighted Cauchy–Schwarz to its cross term:

$$\begin{aligned}\mathcal{E}_{t+1} &\le \sum_i p_i \|a_i^t - \bar{a}^t\|_F^2 + 2\left(\sum_i p_i \|a_i^t - \bar{a}^t\|_F^2\right)^{1/2}\left(\sum_i p_i \|k_i^t - \bar{k}^t\|_F^2\right)^{1/2} \\ &\quad + \sum_i p_i \|k_i^t - \bar{k}^t\|_F^2 \\ &\le \sum_i p_i \|a_i^t\|_F^2 - \|\bar{a}^t\|_F^2 + (2\rho_*\kappa\delta + \kappa^2\delta^2)\mathcal{Q}_t.\end{aligned}$$

Here we used the variance identity and the fact that centering can only reduce a sum of squared norms. Since $\rho_* \le 1$, this proves

$$\mathcal{E}_{t+1} \le \sum_i p_i \|a_i^t\|_F^2 - \|\bar{a}^t\|_F^2 + (2\kappa\delta + \kappa^2\delta^2)\mathcal{Q}_t. \tag{68}$$

Step 3: combine the two inequalities. Use $\mathcal{E}_t = \mathcal{Q}_t - \|\nabla J(V^t)\|_F^2$ and complete the square in $\nabla J(V^t)$ and $\bar{a}^t$:

$$\begin{aligned}&- \|\nabla J(V^t)\|_F^2 + \langle \nabla J(V^t), \bar{a}^t\rangle + \tfrac{1}{2}\left(\sum_i p_i \|a_i^t\|_F^2 - \|\bar{a}^t\|_F^2 - \mathcal{E}_t\right) \\ &\quad = -\tfrac{1}{2}\mathcal{Q}_t + \tfrac{1}{2}\sum_i p_i \|a_i^t\|_F^2 - \tfrac{1}{2}\|\nabla J(V^t) - \bar{a}^t\|_F^2 \\ &\quad \le -\tfrac{1}{2}(1-\rho_*^2)\mathcal{Q}_t.\end{aligned}$$

Since $\mathcal{L}_t = \Delta_t + (s/2)\mathcal{E}_t$, adding (67) to one half of the increment bound from (68) gives

$$\begin{aligned}\frac{\mathcal{L}_{t+1} - \mathcal{L}_t}{s} &\le -\left[\frac{1-\rho_*^2}{2} - \frac{\kappa^2\delta}{2} - \kappa\delta - \frac{\kappa^2\delta^2}{2}\right]\mathcal{Q}_t \\ &\le -\tfrac{1}{4}\mathcal{Q}_t.\end{aligned}$$

Indeed, the bracket is decreasing in $\delta \ge 0$. At $\delta = \delta_0 = 1/10$, $\kappa = 10/9$, and $\rho_* = 1/9$, it equals $17/54 > 1/4$. □

D.3 Parameter Selection and Convergence

To prove Theorem 5.1, choose the constants below. We will verify (62) at every round. Set

$$\begin{aligned}&K_1 := 1 + \Delta_0 + R_0\zeta^2, \\ &G_1 := \sqrt{2LK_1} + \zeta/\sqrt{p_{\min}}, \qquad L_1 := G_1 + 16L(1+2R_0), \\ &c_1 := \min\left\{1, \frac{1}{20L_1}, \frac{1}{4\kappa\sqrt{2LK_1/p_{\min}}}, \frac{p_{\min}}{32\kappa^2 K_1}, \frac{1}{\kappa\sqrt{2560LK_1}}\right\}, \\ &C_2 := 4K_1/c_1, \qquad s = c_1/\sqrt{T}.\end{aligned} \tag{69}$$

Proof of Theorem 5.1. Let $\bar{R} = 2R_0 + 8\kappa^2 K_1 T s$. Initially $C_i^0 = 0$. The variance calculation in Lemma B.2 gives $\mathcal{E}_0 \le R_0\zeta^2$. Therefore

$$0 \le \mathcal{L}_0 = \Delta_0 + \tfrac{s}{2}\mathcal{E}_0 \le \Delta_0 + \tfrac{1}{2}R_0\zeta^2 < K_1,$$

using $s \le c_1 \le 1$. We prove by induction that

$$\mathcal{L}_t \le \mathcal{L}_0, \qquad \sum_{u<t} s\mathcal{Q}_u \le 4K_1, \qquad R_t \le \bar{R}. \tag{70}$$

All three assertions hold at $t = 0$. Suppose they hold at the start of round $t < T$.

Step 1: verify the smoothness condition for this round. Since $\Delta_t \le \mathcal{L}_t < K_1$, Lemma A.1 gives $\max_i \|\nabla F_i(B^t A^t)\|_F \le G_1$. Using the inductive factor bound in the definition of $\ell(V^t)$ gives

$$\ell(V^t) \le G_1 + 16L(1+\bar{R}) = L_1 + 128\kappa^2 L K_1 T s,$$
$$s\ell(V^t) \le c_1 L_1 + 128\kappa^2 L K_1 c_1^2 \le \tfrac{1}{20} + \tfrac{1}{20} = \delta_0.$$

The last line uses the second and fifth restrictions in (69), respectively.

Step 2: verify the local-radius condition. From $\mathcal{L}_t < K_1$ and $\Delta_t \ge 0$, $\mathcal{E}_t \le 2K_1/s$. The global versions of the weight-gradient and factor-gradient bounds also give $\|\nabla J(V^t)\|_F^2 \le 2LK_1R_t$. Thus

$$\mathcal{Q}_t = \|\nabla J(V^t)\|_F^2 + \mathcal{E}_t \le 2LK_1R_t + 2K_1/s.$$

Each $p_i\|g_i^t\|_F^2 \le \mathcal{Q}_t$. Taking square roots and using $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ yields

$$\begin{aligned}\kappa s\|g_i^t\|_F &\le \kappa s\sqrt{\frac{2LK_1}{p_{\min}}}\sqrt{R_t} + \kappa\sqrt{\frac{2K_1 s}{p_{\min}}} \\ &\le \tfrac{1}{4}\sqrt{R_t} + \tfrac{1}{4} \le \tfrac{1}{2}(1 + \|V^t\|_F).\end{aligned}$$

The two terms are controlled by the third and fourth restrictions on c_1 in (69). This verifies both hypotheses of Lemma D.1 before using it.

Step 3: propagate all three induction bounds. The local lemma now places all endpoints in $\mathcal{B}(V^t)$, and Lemma D.2 gives $\mathcal{L}_{t+1} \le \mathcal{L}_t - s\mathcal{Q}_t/4$. Summing this inequality over $u = 0, \dots, t$, and using $\mathcal{L}_{t+1} \ge 0$, proves

$$\mathcal{L}_{t+1} \le \mathcal{L}_0, \qquad \sum_{u \le t} s\mathcal{Q}_u \le 4\mathcal{L}_0 < 4K_1.$$

For each preceding round, the local-radius bound implies $\|\bar{D}^u\|_F \le \kappa s\sqrt{\mathcal{Q}_u}$. Consequently

$$\|V^{t+1} - V^0\|_F \le \sum_{u \le t}\|\bar{D}^u\|_F \le \kappa\sqrt{(t+1)s\sum_{u \le t} s\mathcal{Q}_u},$$
$$R_{t+1} \le 2R_0 + 2\|V^{t+1} - V^0\|_F^2 \le 2R_0 + 8\kappa^2 K_1 T s = \bar{R}.$$

Thus the induction closes without a bounded-factor assumption.

Step 4: obtain the convergence rate. Since $\|\nabla J(V^t)\|_F^2 \le \mathcal{Q}_t$, the gradient sum at $t = T$ and $s = c_1/\sqrt{T}$ give

$$\frac{1}{T}\sum_{t<T}\|\nabla J(V^t)\|_F^2 \le \frac{4K_1}{Ts} = \frac{4K_1/c_1}{\sqrt{T}} = \frac{C_2}{\sqrt{T}}.$$

For $T \ge \lceil (C_2/\epsilon^2)^2 \rceil$, this average is at most ϵ^2. Its minimum is no larger than its average, so one server iterate is ϵ-stationary. The constants in (69) do not depend on K. For $\mu = 0$, take $\eta = s/K = c_1/(K\sqrt{T})$. The local lemma still applies, and every subsequent bound above uses only its conclusions and s, not a positive lower bound on μ. This proves the same full-participation guarantee without a quadratic penalty. □

E Proof of the Partial-Participation Guarantee for FedStabLoRA

We prove Theorem 5.2 with $p_i = 1/N$ and exact local gradients. In each round, S_t is sampled uniformly among the subsets of size M, conditional on the preceding history. The algorithm uses the parameterization in (4).

We temporarily replace each J_i by a function f_i that equals J_i on a ball and has a globally Lipschitz gradient with respect to the factors. A Lyapunov inequality holds for every sequence of participating sets and keeps all local and server iterates in that ball. The auxiliary and original algorithms therefore produce the same iterates. We then take expectations over client sampling to obtain the stated convergence bound.

E.1 Smooth Auxiliary Objectives

For any $R > 0$, define

$$\chi_R(V) = \chi(\|V\|_F^2/R^2), \qquad \chi(u) = \begin{cases} 1, & u \le 1, \\ 1 - \dfrac{(u-1)^2}{3} + \dfrac{2(u-1)^3}{27}, & 1 < u < 4, \\ 0, & u \ge 4. \end{cases} \tag{71}$$

Thus $\chi_R = 1$ on the ball of radius R and $\chi_R = 0$ outside the ball of radius $2R$. Define

$$f_i(V) := J_i(0) + \chi_R(V)[J_i(V) - J_i(0)]. \tag{72}$$

The same cutoff is used for every client. Therefore

$$f(V) := \frac{1}{N}\sum_i f_i(V) = (1 - \chi_R(V))J(0) + \chi_R(V)J(V) \ge F_\star.$$

Subtracting $J_i(0)$ makes the gradient bound independent of additive constants in the client losses.

Lemma E.1 (Smoothness of the auxiliary objectives)**.** *Let $G_3 := \sqrt{2L\Delta_0} + \zeta\sqrt{N} + LR_0/2$. The functions in* (72) *have globally Lipschitz gradients with common constant*

$$\ell_R := 50(G_3 + LR^2).$$

For $\|V\|_F \le R$, we have $f_i(V) = J_i(V)$ and $\nabla f_i(V) = \nabla J_i(V)$.

Proof. **Step 1: bound the original objectives on the radius-$2R$ ball.** Lemma A.1, $p_i = 1/N$, and $\|B^0A^0\|_F \le R_0/2$ imply

$$\begin{aligned} \|\nabla F_i(0)\|_F &\le \|\nabla F_i(B^0A^0)\|_F + L\|B^0A^0\|_F \\ &\le \sqrt{2L\Delta_0} + \zeta\sqrt{N} + LR_0/2 = G_3. \end{aligned}$$

For $\|V\|_F \le 2R$, put $X = BA$. Then $\|X\|_F \le 2R^2$, and weight-gradient smoothness gives $\|\nabla F_i(X)\|_F \le G_3 + 2LR^2$. The chain rule and the two-sided smoothness remainder for F_i give

$$\begin{aligned} \|\nabla J_i(V)\|_F &\le \|V\|_F\|\nabla F_i(X)\|_F \le 2G_3R + 4LR^3, \\ |J_i(V) - J_i(0)| &= |F_i(X) - F_i(0)| \\ &\le G_3\|X\|_F + \tfrac{L}{2}\|X\|_F^2 \le 2G_3R^2 + 2LR^4. \end{aligned}$$

For two points in this ball, the general factor-gradient difference estimate in the proof of Lemma A.2 has coefficient at most $(G_3 + 2LR^2) + 4LR^2 = G_3 + 6LR^2$.

Step 2: bound the cutoff derivatives. On $1 < u < 4$, direct differentiation gives

$$\chi'(u) = \tfrac{2}{9}(u-1)(u-4), \qquad \chi''(u) = \tfrac{1}{9}(4u - 10).$$

Thus $|\chi'| \le 1/2$ and $|\chi''| \le 2/3$ on this interval; $\chi' = 0$ on the two constant pieces and at their endpoints. With $u = \|V\|_F^2/R^2$, the chain rule gives $\nabla\chi_R(V) = 2\chi'(u)V/R^2$, so $\|\nabla\chi_R(V)\|_F \le 2/R$. At points where the second derivative exists, its action on a matrix Z is

$$D(\nabla\chi_R)(V)[Z] = \frac{2\chi'(u)}{R^2}Z + \frac{4\chi''(u)}{R^4}\langle V, Z\rangle V.$$

Its operator norm is at most $R^{-2} + 32/(3R^2) < 12/R^2$ on $\|V\|_F \le 2R$. The first derivatives match at both joins and vanish outside the radius-$2R$ ball. Integrating on line segments therefore gives the global bound $\mathrm{Lip}(\nabla\chi_R) \le 12/R^2$.

Step 3: bound the gradient of the product. Differentiating (72) gives

$$\nabla f_i(V) = \chi_R(V)\nabla J_i(V) + [J_i(V) - J_i(0)]\nabla\chi_R(V).$$

To compare this expression at U, W inside the radius-$2R$ ball, add and subtract $\chi_R(U)\nabla J_i(W)$ in the first term and $[J_i(W) - J_i(0)]\nabla\chi_R(U)$ in the second term. The resulting four differences are

bounded using $0 \le \chi_R \le 1$, Steps 1–2, and the mean-value bound for $J_i(U) - J_i(W)$. Altogether,

$$\begin{aligned}\|\nabla f_i(U) - \nabla f_i(W)\|_F &\le (G_3 + 6LR^2)\|U - W\|_F \\ &\quad + \frac{4}{R}(2G_3R + 4LR^3)\|U - W\|_F \\ &\quad + \frac{12}{R^2}(2G_3R^2 + 2LR^4)\|U - W\|_F \\ &= (33G_3 + 46LR^2)\|U - W\|_F \le \ell_R\|U - W\|_F.\end{aligned}$$

This argument does not assume that F_i has a continuous Hessian.

Step 4: extend the bound to the whole factor space. For $\|V\|_F \ge 2R$, both χ_R and $\nabla\chi_R$ vanish, so $\nabla f_i(V) = 0$, continuously including the boundary. If one point is outside the ball and one is inside, apply Step 3 between the inside point and the first boundary point on their connecting segment. The outside gradient and boundary gradient are both zero, and that subsegment is no longer than the full segment. If both points are outside, both gradients are zero. This proves global ℓ_R-Lipschitz continuity. On $\|V\|_F \le R$, the identities $\chi_R = 1$ and $\nabla\chi_R = 0$ prove equality of both the functions and their gradients with J_i. □

E.2 State Identity and Tracking Error

For the next two subsections, run FedStabLoRA with f_i in place of J_i, using the same notation for its iterates and states. We will show in Section E.4 that this auxiliary run is identical to the original run. Set $h := Ns/M$ and, for now, suppose $h\ell_R \le \delta_0$.

Let $\bar{D}^t := M^{-1}\sum_{i\in S_t} D_i^t$. The updates are $V^{t+1} = V^t + \bar{D}^t$, $\bar{Y}^{t+1} = \bar{Y}^t - \frac{M}{Ns}\bar{D}^t$, and $Y_i^{t+1} = Y_i^t - D_i^t/s$ for selected clients. Other client states do not change. These updates preserve $\bar{Y}^t = N^{-1}\sum_i Y_i^t$ and give the exact identity

$$V^t + h\bar{Y}^t = V^0.$$

To verify this, use $h = Ns/M$ in the two server updates:

$$\begin{aligned}V^{t+1} + h\bar{Y}^{t+1} &= V^t + \bar{D}^t + h\bar{Y}^t - \frac{hM}{Ns}\bar{D}^t \\ &= V^t + h\bar{Y}^t.\end{aligned}$$

Induction from $\bar{Y}^0 = 0$ proves the identity for every sequence of participating sets.

For the analysis, define

$$Z_i^t := V^0 - hY_i^t, \qquad \psi_i(V) := \tfrac{1}{2}\|V\|_F^2 - hf_i(V).$$

The state identity gives $N^{-1}\sum_i Z_i^t = V^0 - h\bar{Y}^t = V^t$. With $g_i^t := \nabla f_i(V^t) - C_i^t$ and $\mathcal{Q}_t := N^{-1}\sum_i \|g_i^t\|_F^2$, expand

$$\begin{aligned}Z_i^t - \nabla\psi_i(V^t) &= V^0 - hY_i^t - V^t + h\nabla f_i(V^t) \\ &= h[\nabla f_i(V^t) - Y_i^t + \bar{Y}^t] = hg_i^t.\end{aligned}$$

Thus

$$Z_i^t - \nabla\psi_i(V^t) = hg_i^t. \tag{73}$$

Since $s = (M/N)h \le h$ and $h\ell_R \le \delta_0$, the global version of Lemma D.1 applies. Its endpoint identity is $D_i^t = -sg_i^t + sa_i^t$, with $\|a_i^t\|_F \le \rho_*\|g_i^t\|_F$. For a selected client, substitute this into $Z_i^{t+1} = Z_i^t + (h/s)D_i^t$ and use (73). This gives

$$Z_i^{t+1} = Z_i^t + (h/s)D_i^t = \nabla\psi_i(V^t) + ha_i^t, \qquad \|a_i^t\|_F \le \rho_*\|g_i^t\|_F. \tag{74}$$

Unselected states Z_i^t remain unchanged. By (73) and (74), updating a selected state makes its distance to $\nabla\psi_i(V^t)$ at most ρ_* times the previous distance.

E.3 Pathwise Lyapunov Descent

The smoothness remainder of f_i satisfies

$$|f_i(U) - f_i(V) - \langle \nabla f_i(V), U - V\rangle| \le \tfrac{1}{2}\ell_R \|U - V\|_F^2.$$

Subtract h times this remainder from the exact quadratic remainder $\|U - V\|_F^2/2$. Since $h\ell_R \le \delta_0 < 1$, we obtain, for all U, V,

$$\frac{1-\delta_0}{2}\|U - V\|_F^2 \le \psi_i(U) - \psi_i(V) - \langle \nabla\psi_i(V), U - V\rangle \le \frac{1+\delta_0}{2}\|U - V\|_F^2. \tag{75}$$

Let $\psi_i^*(z) := \sup_U\{\langle z, U\rangle - \psi_i(U)\}$ and define

$$\mathcal{G}_i(z, V) := \psi_i^*(z) + \psi_i(V) - \langle z, V\rangle.$$

Inequality (75) makes ψ_i $(1-\delta_0)$-strongly convex, whether or not f_i is convex. In particular, the supremum defining ψ_i^* is finite: its quadratic lower bound makes $\langle z, U\rangle - \psi_i(U)$ tend to $-\infty$ as $\|U\|_F \to \infty$.

For fixed z, V, set $Z = U - V$ inside this supremum. By the definition of $\mathcal{G}_i$,

$$\mathcal{G}_i(z, V) = \sup_Z\big\{\langle z - \nabla\psi_i(V), Z\rangle - [\psi_i(V + Z) - \psi_i(V) - \langle \nabla\psi_i(V), Z\rangle]\big\}.$$

For any matrix b and scalar $c > 0$,

$$\sup_Z\{\langle b, Z\rangle - \tfrac{c}{2}\|Z\|_F^2\} = \frac{\|b\|_F^2}{2c},$$

as follows by completing the square, with maximizer $Z = b/c$. Apply this identity with $c = 1 + \delta_0$ for the lower bound and $c = 1 - \delta_0$ for the upper bound. Then (75) yields

$$\frac{\|z - \nabla\psi_i(V)\|_F^2}{2(1+\delta_0)} \le \mathcal{G}_i(z, V) \le \frac{\|z - \nabla\psi_i(V)\|_F^2}{2(1-\delta_0)}. \tag{76}$$

Both bounds imply that $\mathcal{G}_i(z, V) = 0$ exactly when $z = \nabla\psi_i(V)$. No convexity assumption on f_i has been added.

Define the Lyapunov function

$$\mathcal{P}_t := \frac{1}{h}\left[\frac{1}{N}\sum_i \psi_i^*(Z_i^t) - \frac{1}{2}\|V^t\|_F^2\right] - F_\star. \tag{77}$$

Expand the average gap and use $N^{-1}\sum_i Z_i^t = V^t$:

$$\begin{aligned}\frac{1}{N}\sum_i \mathcal{G}_i(Z_i^t, V^t) &= \frac{1}{N}\sum_i \psi_i^*(Z_i^t) + \tfrac{1}{2}\|V^t\|_F^2 - hf(V^t) - \left\langle \frac{1}{N}\sum_i Z_i^t, V^t\right\rangle \\ &= \frac{1}{N}\sum_i \psi_i^*(Z_i^t) - \tfrac{1}{2}\|V^t\|_F^2 - hf(V^t).\end{aligned}$$

Rearrange and apply the lower gap bound together with (73). We obtain

$$\mathcal{P}_t = f(V^t) - F_\star + \frac{1}{hN}\sum_i \mathcal{G}_i(Z_i^t, V^t) \ge f(V^t) - F_\star + \frac{h}{2(1+\delta_0)}\mathcal{Q}_t \ge 0. \tag{78}$$

Thus an upper bound on $\mathcal{P}_t$ bounds both the objective gap and $\mathcal{Q}_t$.

Lemma E.2 (Descent for every participating set)**.** *For every S_t of size M,*

$$\mathcal{P}_{t+1} \le \mathcal{P}_t - \frac{h}{4N}\sum_{i\in S_t}\|g_i^t\|_F^2 - \frac{1}{2h}\|V^{t+1} - V^t\|_F^2. \tag{79}$$

Proof. **Step 1: compare the old and new states at V^t.** For each selected client, (76) and (74) give

$$\mathcal{G}_i(Z_i^{t+1}, V^t) \le \frac{h^2\|a_i^t\|_F^2}{2(1-\delta_0)} \le \frac{\rho_*^2 h^2\|g_i^t\|_F^2}{2(1-\delta_0)}.$$

The old-state lower bound from (73) is $\mathcal{G}_i(Z_i^t, V^t) \ge h^2\|g_i^t\|_F^2/[2(1+\delta_0)]$. Subtract the latter from the former:

$$\begin{aligned}\mathcal{G}_i(Z_i^{t+1}, V^t) - \mathcal{G}_i(Z_i^t, V^t) &\le -\left[\frac{1}{2(1+\delta_0)} - \frac{\rho_*^2}{2(1-\delta_0)}\right] h^2\|g_i^t\|_F^2 \\ &\le -\tfrac{1}{4}h^2\|g_i^t\|_F^2.\end{aligned}$$

At $\delta_0 = 1/10$ and $\rho_* = 1/9$, the bracket equals $5/11 - 5/729 > 1/4$. Unselected clients have $Z_i^{t+1} = Z_i^t$ and hence zero change when the comparison point is kept at V^t.

Step 2: account exactly for the server movement. The two state means equal the corresponding server iterates, so $N^{-1}\sum_i(Z_i^{t+1} - Z_i^t) = V^{t+1} - V^t$. At the fixed comparison point V^t, the gap definition gives

$$\frac{1}{N}\sum_i[\mathcal{G}_i(Z_i^{t+1}, V^t) - \mathcal{G}_i(Z_i^t, V^t)] = \frac{1}{N}\sum_i[\psi_i^*(Z_i^{t+1}) - \psi_i^*(Z_i^t)] - \langle V^{t+1} - V^t, V^t\rangle.$$

Meanwhile, expand the quadratic term in (77):

$$\begin{aligned}h(\mathcal{P}_{t+1} - \mathcal{P}_t) &= \frac{1}{N}\sum_i[\psi_i^*(Z_i^{t+1}) - \psi_i^*(Z_i^t)] - \langle V^t, V^{t+1} - V^t\rangle - \tfrac{1}{2}\|V^{t+1} - V^t\|_F^2 \\ &= \frac{1}{N}\sum_{i\in S_t}[\mathcal{G}_i(Z_i^{t+1}, V^t) - \mathcal{G}_i(Z_i^t, V^t)] - \tfrac{1}{2}\|V^{t+1} - V^t\|_F^2.\end{aligned}$$

The inner products cancel exactly; no expectation is used. Insert Step 1 and divide by $h > 0$ to obtain (79). □

Before taking any expectations, (79) and (78) give, on every participation history,

$$\begin{aligned}&\mathcal{P}_t \le \mathcal{P}_0, \qquad \mathcal{Q}_t \le 2(1+\delta_0)\mathcal{P}_0/h, \\ &\sum_{t<T}\|V^{t+1} - V^t\|_F^2 \le 2h\mathcal{P}_0, \quad \|V^t\|_F^2 \le 2R_0 + 4Th\mathcal{P}_0 \quad (t \le T).\end{aligned} \tag{80}$$

Indeed, nonnegativity of $\mathcal{P}_t$ and telescoping (79) give $\sum_{u<t}\|V^{u+1} - V^u\|_F^2 \le 2h\mathcal{P}_0$. Consequently,

$$\begin{aligned}\|V^t - V^0\|_F^2 \le t\sum_{u<t}\|V^{u+1} - V^u\|_F^2 &\le 2Th\mathcal{P}_0, \\ \|V^t\|_F^2 \le 2R_0 + 2\|V^t - V^0\|_F^2 &\le 2R_0 + 4Th\mathcal{P}_0.\end{aligned}$$

The bound on $\mathcal{Q}_t$ follows by dropping the nonnegative objective gap in (78) and using $\mathcal{P}_t \le \mathcal{P}_0$. These bounds hold for every sequence of sets S_t of size M; uniform sampling will be used only when taking expectations.

E.4 Containment and Equivalence to the Original Algorithm

We now specify R and the constants c_3, C_4 from Theorem 5.2. Define

$$\begin{aligned}&K_3 := 1 + \Delta_0 + R_0(\zeta^2 + 2L\Delta_0), \\ &\widetilde{L}_3 := 50\left(\sqrt{2L\Delta_0} + \zeta + \frac{LR_0}{2}\right) + 400L(R_0+1), \\ &c_3 := \min\left\{1, \frac{1}{20\widetilde{L}_3}, \frac{1}{\sqrt{16000LK_3}}\right\}, \qquad C_4 := 4K_3/c_3, \\ &h = \frac{c_3}{\sqrt{NT}}, \qquad s = \frac{M}{N}h, \qquad R^2 = 8[R_0 + 1 + K_3h(T+N)].\end{aligned} \tag{81}$$

The constants c_3 and C_4 are independent of N, M, K, and T. **Step 1: verify global smoothness at the selected parameters.** The choices in (81) fix the auxiliary objectives before their run begins. Since $N \geq 1$, $G_3 \leq (\sqrt{2L\Delta_0} + \zeta + LR_0/2)\sqrt{N}$. Substitution of R^2 into ℓ_R gives

$$\begin{aligned} h\ell_R &\leq \widetilde{L}_3\sqrt{N}\,h + 400LK_3h^2(T+N) \\ &= \frac{\widetilde{L}_3 c_3}{\sqrt{T}} + 400LK_3c_3^2\left(\frac{1}{N}+\frac{1}{T}\right) \\ &\leq \widetilde{L}_3 c_3 + 800LK_3c_3^2 \leq \delta_0. \end{aligned}$$

Here $N, T \geq 1$, so $1/N + 1/T \leq 2$. The two last terms are at most $\delta_0/2$ by the two corresponding restrictions on c_3. Thus the global auxiliary analysis applies without assuming that its iterates remain in the agreement ball.

Step 2: bound the initial potential. The definition of R gives $R^2 \geq 8(R_0+1) > R_0$. Thus V^0 is in the agreement ball and $f_i(V^0) = J_i(V^0)$. At initialization $Z_i^0 = V^0$ and $C_i^0 = 0$. The initial-gradient bound from Lemmas A.1 and A.2 is $N^{-1}\sum_i \|\nabla J_i(V^0)\|_F^2 \leq R_0(\zeta^2 + 2L\Delta_0)$. Therefore

$$\mathcal{P}_0 \leq \Delta_0 + \frac{h}{2(1-\delta_0)}\frac{1}{N}\sum_i \|\nabla J_i(V^0)\|_F^2 < K_3.$$

We used $h \leq 1$ and $[2(1-\delta_0)]^{-1} < 1$.

Step 3: place all local and server iterates in the ball. For every client i, $\|g_i^t\|_F^2 \leq N\mathcal{Q}_t$. Apply the global local-update bound and (80), followed by $\mathcal{P}_0 < K_3$. For every local step at round $t < T$, this gives

$$\begin{aligned} \|V_{i,e}^t\|_F^2 &\leq 2\|V^t\|_F^2 + 2\kappa^2 s^2 \|g_i^t\|_F^2 \\ &\leq 4R_0 + 8ThK_3 + 4\kappa^2(1+\delta_0)\frac{M^2}{N}hK_3 \\ &\leq 4R_0 + 8ThK_3 + 6NhK_3 < R^2. \end{aligned}$$

Here $\|g_i^t\|_F^2 \leq N\mathcal{Q}_t$ and $4\kappa^2(1+\delta_0) = 440/81 < 6$. The server iterates also satisfy $\|V^t\|_F^2 \leq 2R_0 + 4ThK_3 < R^2$ for $t \leq T$. The same bound holds for the local iterates that any client would produce if selected at round t.

Step 4: identify the auxiliary run with the original run. Couple the two runs by using the same initialization and the same sets S_t. They have identical server iterates and correction states at $t = 0$. Suppose these agree at the start of round t. For each selected client, both runs start at the same factors and use the same centered correction. At each local step, the auxiliary iterate lies inside the agreement ball by Step 3, where $\nabla f_i = \nabla J_i$. The two local updates are therefore equal. Induction over the local steps gives equal endpoints and equal displacements D_i^t. The server and state updates then agree as well, completing the induction over rounds. The cutoff is only a proof device and does not modify Algorithm 1.

E.5 Expected Convergence

Proof of Theorem 5.2. Let $\mathcal{F}_t$ be the history before sampling S_t, and write $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_t]$. All states and the local update that any client would produce in round t are determined by $\mathcal{F}_t$, since gradients are exact. The preceding equivalence argument transfers (79) to the original run for every history.

Step 1: average over the clients selected in one round. Uniform sampling of a size-M subset gives $\mathbb{E}_t[\mathbf{1}\{i \in S_t\}] = M/N$. As g_i^t is $\mathcal{F}_t$-measurable,

$$\begin{aligned} \mathbb{E}_t\left[\sum_{i\in S_t}\|g_i^t\|_F^2\right] &= \sum_i \|g_i^t\|_F^2 \mathbb{E}_t[\mathbf{1}\{i \in S_t\}] \\ &= \frac{M}{N}\sum_i \|g_i^t\|_F^2 = M\mathcal{Q}_t. \end{aligned}$$

Only these inclusion probabilities are needed; independence of the indicators within a round is not assumed. Drop the nonpositive movement term in (79), take conditional expectation, and use $hM/N = s$:

$$\mathbb{E}_t \mathcal{P}_{t+1} \le \mathcal{P}_t - \tfrac{s}{4}\mathcal{Q}_t.$$

Centering gives $N^{-1}\sum_i C_i^t = 0$. On the agreement ball, $N^{-1}\sum_i g_i^t = \nabla J(V^t)$, so Jensen's inequality implies $\|\nabla J(V^t)\|_F^2 \le \mathcal{Q}_t$.

Step 2: telescope unconditional expectations. The pathwise bounds give $0 \le \mathcal{P}_t \le \mathcal{P}_0$ and $\mathcal{Q}_t \le 2(1+\delta_0)\mathcal{P}_0/h$, so all quantities above are integrable. Taking full expectation and using the tower property yields

$$\tfrac{s}{4}\mathbb{E}\|\nabla J(V^t)\|_F^2 \le \mathbb{E}\mathcal{P}_t - \mathbb{E}\mathcal{P}_{t+1}.$$

Sum over $t = 0, \dots, T-1$ and use $\mathcal{P}_T \ge 0$:

$$\frac{s}{4}\sum_{t<T} \mathbb{E}\|\nabla J(V^t)\|_F^2 \le \mathcal{P}_0 - \mathbb{E}\mathcal{P}_T \le \mathcal{P}_0 < K_3. \tag{82}$$

Step 3: convert the bound to communication complexity. Divide by $sT/4$ and substitute $s = c_3 M/(N^{3/2}\sqrt{T})$:

$$\frac{1}{T}\sum_{t<T} \mathbb{E}\|\nabla J(V^t)\|_F^2 \le \frac{4K_3}{Ts} = \frac{C_4 N^{3/2}}{M\sqrt{T}}.$$

For $T \ge \lceil C_4^2 N^3/(M^2\epsilon^4)\rceil$, the right-hand side is at most ϵ^2, and hence

$$\min_{0\le t<T} \mathbb{E}\|\nabla J(V^t)\|_F^2 \le \frac{1}{T}\sum_{t<T} \mathbb{E}\|\nabla J(V^t)\|_F^2 \le \epsilon^2.$$

The minimum is taken after expectation, exactly as in the theorem. This does not assert a last-iterate or pathwise stationarity guarantee. For $\mu = 0$, choose $\eta = s/K = c_3 M/(KN^{3/2}\sqrt{T})$. The local lemma remains valid for the auxiliary objectives, and the state identity, Lyapunov descent, containment, and expectation arguments are unchanged. Thus the same partial-participation bound holds without a quadratic penalty. □

Table 3: Dataset statistics and split sizes. *Train* is partitioned across clients, *Tune* is used for hyperparameter and checkpoint selection, and *Eval* is used only for final reporting. For GLUE, *Eval* is the public validation set; for AG News, CIFAR-100, and DomainNet-100, it is the official test set.

Dataset	Classes	Train	Tune	Eval
MNLI-m	3	353,238	39,248	9,815
QQP	2	327,148	36,349	40,430
QNLI	2	94,268	10,475	5,463
SST-2	2	60,276	6,697	872
AG News	4	107,891	11,988	7,600
CIFAR-100	100	45,000	5,000	10,000
DomainNet-100	100	100,031	11,101	47,937

F Experimental Details and Additional Results

This appendix provides implementation and evaluation details for the experiments in Section 6, together with additional results referenced in the main text.

F.1 Datasets and Evaluation Protocol

Each benchmark uses three non-overlapping subsets. *Train* is the data pool partitioned across clients and used for local training. *Tune* is held out from local training and is used for hyperparameter and checkpoint selection. *Eval* is never used for selection. For each run, we select the communication round with the best metric on *Tune* and report the corresponding metric on *Eval*. All classification benchmarks, including QQP, are evaluated by accuracy.

For the text benchmarks, we remove duplicate examples and any training examples that overlap with *Eval*, then reserve a label-stratified 10% of the remaining training data as *Tune*. The same data split is used across methods. For the GLUE tasks, *Eval* is the public validation set because test labels are not released; MNLI-m uses the matched validation set. AG News and CIFAR-100 use their official test sets. For DomainNet-100, we use the first 100 classes from all six domains and combine the corresponding official test lists for *Eval*. Table 3 reports the resulting split sizes.

Text inputs are truncated or padded to 128 tokens. CIFAR-100 and DomainNet images are resized to 224×224. The Qwen2.5 scaling experiments use the same MNLI-m examples and split as the RoBERTa-base experiments, retokenized with the corresponding Qwen tokenizer.

F.2 Client Partitions and Randomness

We keep the number of training examples the same across clients and vary only their label distributions. For a training pool with M examples and N clients, each client receives

$$n = \left\lfloor \frac{M}{N} \right\rfloor$$

examples. If M is not divisible by N, we remove at most $N - 1$ examples, with the removals distributed across classes according to their frequencies.

For a non-IID split, for each class c we draw

$$p_c \sim \mathrm{Dirichlet}(\alpha \mathbf{1}_N).$$

We use p_c to determine how examples of class c are distributed across clients, while enforcing two constraints: every client receives exactly n examples, and every retained training example is assigned once. Hence α changes only the label distribution across clients, not the number of examples stored by each client. The main experiments use $\alpha = 0.1$. The heterogeneity study also uses $\alpha = 1.0$ and IID. For IID, we randomly permute the retained training pool and split it into N equal-size shards.

Unless otherwise stated, we average results over three seeds. A seed determines the model initialization, dropout, minibatch order, and client partition. For a fixed seed, all compared methods use the

Table 4: Training configurations for the language, vision, and model-scaling experiments. All reported experiments use constant-step SGD. The local-epoch budget E and corresponding number K of local SGD steps are reported in Table 5.

Setting	Language (RoBERTa-base)	Vision (ViT-S/16)	Scaling (Qwen2.5)
Optimizer	SGD	SGD	SGD
Momentum	0	0	0
Weight decay	0	0	0
Batch size	128	64	16, gradient accumulation 8
Learning-rate schedule	constant	constant	constant
Gradient clipping	none	none	none
LoRA rank r	4	4	8
LoRA scaling α_{LoRA}	8	8	16
LoRA dropout	0.1	0.05	0.1
LoRA attention targets	Q, V	Q, K, V	Q, V
Number of clients	10	10	10
Communication rounds	50	50	50

same client datasets. In the partial-participation experiments, clients are sampled uniformly without replacement in each round, and all methods use the same sampled clients for a fixed seed and round.

All clients have the same number of training examples. Therefore sample-count-weighted aggregation and uniform aggregation are identical, and we use uniform client weights throughout the experiments.

F.3 Backbones, LoRA Configuration, and Optimization

Table 4 summarizes the common training configurations for the language, vision, and model-scaling experiments. For RoBERTa-base and Qwen2.5, LoRA is applied to the Q and V projections in each attention layer. For ViT-S/16, LoRA is applied to the Q, K, and V projections. In all three cases, the task-specific classification head is trained densely and federated together with the LoRA parameters. We use the standard LoRA initialization with the B factor initialized to zero.

Local epochs and optimizer steps. In the experimental section, E denotes the local-data budget in equivalent epochs. For a client with n training examples and effective batch size B_{eff}, we implement this budget with a fixed number

$$K = \left\lceil \frac{En}{B_{\mathrm{eff}}} \right\rceil$$

of local SGD updates per participating client and round. Thus E is the local-training budget, while K is the exact optimizer-step count used by the implementation and by the theoretical sections. Within a setting, every client uses the same K. Minibatches are reshuffled and reused as needed until exactly K updates have been completed.

The main benchmarks use $E = 2$. The local-training study varies $E \in \{1, 2, 4\}$. The partial-participation and Qwen2.5 scaling studies use $E = 1$. Table 5 reports both E and the resulting K for every setting.

Local SGD update. All experiments use plain SGD with zero momentum, zero weight decay, no gradient clipping, and a constant local learning rate. Let θ denote the full trainable parameter vector used in the implementation, including the LoRA factors and task-specific classification head. At local step k of round t, client i draws a minibatch $\mathcal{B}_{i,k}^t$ and computes its minibatch gradient $\widehat{g}_i(\theta_{i,k}^t; \mathcal{B}_{i,k}^t)$. FedAvg-LoRA then performs the plain-SGD step

$$\theta_{i,k+1}^t = \theta_{i,k}^t - \eta\, \widehat{g}_i(\theta_{i,k}^t; \mathcal{B}_{i,k}^t).$$

FedStabLoRA modifies this minibatch gradient before applying the same SGD step:

$$\theta_{i,k+1}^t = \theta_{i,k}^t - \eta\big[\widehat{g}_i(\theta_{i,k}^t; \mathcal{B}_{i,k}^t) - C_i^t + \mu(\theta_{i,k}^t - \theta^t)\big]\,.$$

The centered correction C_i^t is held fixed during all K local steps in the round. After the final step, the client returns $D_i^t = \theta_{i,K}^t - \theta^t$; the server averages the returned displacements, and the client

Table 5: Client sizes and local-update budgets. E denotes the equivalent number of local data passes, and $K = \lceil En/B_{\text{eff}} \rceil$ is the fixed number of SGD steps performed by each participating client. The theory uses the same K as its local-step count.

Setting	N	n	B_{eff}	E	K
MNLI-m (main/heterogeneity)	10	35,323	128	2	552
QQP	10	32,714	128	2	512
QNLI	10	9,426	128	2	148
SST-2	10	6,027	128	2	95
AG News	10	10,789	128	2	169
CIFAR-100	10	4,500	64	2	141
DomainNet-100	10	10,003	64	2	313
MNLI-m (local training, $E = 1$)	10	35,323	128	1	276
MNLI-m (local training, $E = 2$)	10	35,323	128	2	552
MNLI-m (local training, $E = 4$)	10	35,323	128	4	1104
MNLI-m (partial participation)	50	7,064	128	1	56
MNLI-m (Qwen2.5 scaling)	10	35,323	128[†]	1	276

[†] Qwen2.5 uses minibatch size 16 with gradient accumulation over 8 minibatches, giving $B_{\text{eff}} = 128$ examples per optimizer step.

correction state is updated from D_i^t using the scale s below. FedProx, FedDyn, and SCAFFOLD use the same minibatches, plain-SGD optimizer, and local-step budget. Each method applies its prescribed local regularization or correction, together with its method-specific auxiliary-state and server updates where applicable. In particular, FedDyn also applies its dynamic server-side correction after aggregating the returned client models. This protocol makes K the exact number of optimizer updates performed per participating client and communication round.

Correction scale. Every run uses a constant local learning rate η. For FedStabLoRA, we set

$$s = \eta \sum_{k=0}^{K-1} (1 - \eta\mu)^k,$$

where K is the fixed number of local SGD updates corresponding to the reported epoch budget E, and η and μ are the selected hyperparameters. Hence s is fixed across clients and communication rounds within a run and is not tuned separately.

The vision runs use random crop with padding 4 and horizontal flip during local training. All experiments use bfloat16 autocast for forward and backward computation, while the trainable master parameters, federated parameter vectors, and server aggregation remain in fp32. The RoBERTa-base configuration has $P = 740{,}355$ trainable parameters; we use this value of P to report communication and state costs in Appendix F.11.

F.4 Federated Methods and Hyperparameter Selection

Common protocol. Whenever a method aggregates client quantities by averaging, we use uniform client weights. Because every client has the same number of training examples, uniform weighting is equivalent to weighting by local sample count. Each method then applies its method-specific client-state and server-state updates according to its original algorithm. Within each benchmark, we keep the backbone, optimizer family, batch size, local-epoch budget E, the corresponding fixed number K of local SGD updates, and communication-round budget fixed across methods. The learning rate is selected separately for each method.

Baseline implementation. FedDyn follows the dynamic-regularization update of Acar et al. (2021). Let θ_i^{t+1} denote the model returned by participating client i after its local updates. The server first forms

$$\bar{\theta}^{t+1} = \frac{1}{|\mathcal{S}_t|} \sum_{i \in \mathcal{S}_t} \theta_i^{t+1},$$

Table 6: Method-specific hyperparameter search spaces used in the main experiments. The local learning rate for every method is selected from $\{0.3, 0.1, 0.03, 0.01, 0.003\}$.

Method	Additional search space
FedProx	$\mu \in \{0.1,\ 0.01,\ 0.001,\ 0.0001\}$
FedDyn	$\alpha \in \{0.1,\ 0.01,\ 0.001,\ 0.0001\}$
SCAFFOLD	Option II
FedStabLoRA	$\mu \in \{0.3,\ 0.1,\ 0.03,\ 0.01\}$

Table 7: Selected learning rates for the main experiments. Each method is tuned independently on *Tune*.

Dataset	FedAvg	FedProx	FedDyn	SCAFFOLD	FedEx-LoRA	FFA-LoRA	FedStabLoRA
MNLI-m	0.03	0.1	0.1	0.01	0.03	0.03	0.3
QQP	0.03	0.03	0.1	0.01	0.01	0.01	0.1
QNLI	0.1	0.03	0.1	0.01	0.03	0.03	0.3
SST-2	0.1	0.03	0.1	0.01	0.03	0.03	0.3
AG News	0.1	0.03	0.3	0.01	0.03	0.01	0.1
CIFAR-100	0.03	0.03	0.1	0.003	0.01	0.03	0.1
DomainNet-100	0.03	0.1	0.1	0.003	0.03	0.03	0.1

updates the FedDyn server state as

$$h^{t+1} = h^t - \frac{\alpha}{N} \sum_{i \in \mathcal{S}_t} (\theta_i^{t+1} - \theta^t),$$

and then sets

$$\theta^{t+1} = \bar{\theta}^{t+1} - \frac{1}{\alpha} h^{t+1}.$$

Thus, the FedDyn global model is not simply the average of the returned client models; the averaged model is followed by the method-specific server correction. FedEx-LoRA computes the residual induced by separately aggregating the LoRA factors and adds it to the frozen base weights. This residual must be available to participating clients in the next round. Following the exact communication protocol of Singhal et al. (2025), we account for this server-to-client residual transmission separately rather than treating the server-side weight update as free communication. FFA-LoRA freezes the A factor and trains only B. For SCAFFOLD, we use Option II throughout the experiments.

Hyperparameter selection. All hyperparameter selection is performed on *Tune*. For each method and benchmark, we tune the local learning rate independently and tune any additional method-specific coefficient on the same split. SCAFFOLD uses Option II in all experiments. The selected configuration is then run with three seeds, and we report mean $\pm$ standard deviation on *Eval*.

Hyperparameter search spaces. For every method and benchmark in the main comparison, we tune the local SGD learning rate over

$$\eta \in \{0.3,\ 0.1,\ 0.03,\ 0.01,\ 0.003\}.$$

These values are larger than learning rates commonly used for full-parameter fine-tuning because our experiments optimize only the LoRA factors and the task head if necessary with constant-step SGD. In preliminary tuning, the useful learning-rate range for these LoRA runs was concentrated in this larger range, so we use the same grid for all compared methods.

FedAvg, FedEx-LoRA, and FFA-LoRA have no additional method-specific coefficient. SCAFFOLD uses Option II throughout. For the remaining methods, we use the search spaces in Table 6. For FedStabLoRA, the correction scale s is computed from (K, η, μ) using (4) and is not tuned independently.

Selected learning rates. Table 7 records the learning rates selected for the main experiments. Each entry is selected independently on *Tune* for the corresponding method and benchmark.

Table 8: Selected method-specific hyperparameters for the main experiments.

Dataset	FedProx μ	FedDyn α	FedStabLoRA μ
MNLI-m	0.01	0.01	0.03
QQP	0.01	0.01	0.1
QNLI	0.1	0.001	0.1
SST-2	0.01	0.01	0.3
AG News	0.1	0.1	0.1
CIFAR-100	0.01	0.01	0.03
DomainNet-100	0.01	0.001	0.1

Table 9: Configurations for the LoRA component ablation in Table 2. The ablation is centered at the selected MNLI-m FedStabLoRA configuration, $\eta = 0.3$ and $\mu = 0.03$. The correction-only variant diverges at $\eta = 0.3$ and is therefore retuned on *Tune*.

Variant	η	μ	s	Selection
Base	0.3	0	–	fixed
+Penalty	0.3	0.03	–	fixed
+Correction	0.01	0	$K\eta$	LR retuned
+Both ($s = K\eta$)	0.3	0.03	165.6	fixed
FedStabLoRA	0.3	0.03	33.11	selected on *Tune*

Selected method-specific hyperparameters. After selecting the learning rate, we select the remaining method-specific hyperparameters on *Tune*. Table 8 lists the resulting configurations. FedAvg, FedEx-LoRA, and FFA-LoRA have no additional algorithmic coefficient.

FedStabLoRA implementation and ablation. FedStabLoRA uses the correction scale in (4), except where the scale is explicitly changed in the ablations. At local step k, client i applies SGD to

$$\nabla J_i(V_{i,k}) - C_i + \mu(V_{i,k} - V)$$

with the constant learning rate η. The value of s is computed once from (K, η, μ) as described above and then kept fixed for the run. After the learning rate is selected, μ is the only additional continuous hyperparameter. In the component ablation, we remove either the quadratic penalty or the drift correction. For the LoRA correction-only variant, training diverges at the FedStabLoRA learning rate $\eta = 0.3$, so we retune its learning rate on *Tune*, as noted in Table 2. Its scale $s = K\eta$ is computed using the retuned learning rate. For the two variants that retain both terms, the selected learning rate and μ are unchanged; only the correction scale differs.

F.5 Software and Hardware

Experiments use PyTorch 2.13.0 with CUDA 13.0, Transformers 5.15.0, and PEFT 0.20.0. Runs are executed on NVIDIA H200 (141 GiB), H100 (80 GiB), and L40S (48 GiB) GPUs. For runtime profiling, all compared methods are run under the same hardware setting so that the reported differences are not confounded by GPU type.

F.6 Ablation Configurations

Table 2 is a controlled component ablation around the selected FedStabLoRA configuration, rather than a retuned comparison of the corresponding baseline methods in Table 1. Unless otherwise stated, the ablation variants use the same client partitions, random seeds, local-update budget, and communication budget as FedStabLoRA, while individual components are removed or modified. The correction-only variant is the exception: because it diverges at the FedStabLoRA learning rate, its learning rate is retuned on the tuning split.

Under matched hyperparameters, initialization, client partitions, and minibatch order, the Base and +Penalty variants correspond to FedAvg and FedProx, respectively. Under full participation, the +Correction variant with $\mu = 0$ and $s = K\eta$ corresponds to SCAFFOLD Option II; see Appendix G.

Table 10: MNLI-m accuracy (%) under different levels of label heterogeneity. All runs use 10 clients, full participation, and $E = 2$. Results are mean $\pm$ standard deviation over three seeds.

Method	Dirichlet $\alpha = 0.1$	Dirichlet $\alpha = 1.0$	IID
FedAvg	76.27±2.36	80.27±1.36	85.38±0.06
FedDyn	82.71±1.02	83.70±0.83	86.81±0.10
SCAFFOLD	81.25±1.64	83.25±0.71	86.95±0.12
FedStabLoRA	**85.07±0.65**	86.38±0.99	87.09±0.09

Table 11: MNLI-m accuracy (%) as the number of local epochs varies under Dirichlet label skew ($\alpha = 0.1$). All runs use 10 clients and full participation. Results are mean $\pm$ standard deviation over three seeds.

Method	$E = 1$	$E = 2$	$E = 4$
FedAvg	80.85±1.88	76.27±2.36	71.09±3.47
FedDyn	82.69±0.38	82.71±1.02	81.21±1.77
SCAFFOLD	82.95±0.50	81.25±1.64	80.20±0.92
FedStabLoRA	85.03±0.55	**85.07±0.65**	84.79±0.99

The different values reported in Tables 1 and 2 therefore reflect different hyperparameter-selection protocols rather than different algorithm definitions.

F.7 Data Heterogeneity and Local Training

Experimental setup. Both studies use MNLI-m with RoBERTa-base, 10 clients, full participation, and 50 communication rounds. For the heterogeneity study, we fix $E = 2$ and vary the client partition over $\alpha \in \{0.1, 1.0\}$ and IID. For the local-step study, we fix $\alpha = 0.1$ and vary $E \in \{1, 2, 4\}$. We compare FedAvg, FedDyn, SCAFFOLD, and FedStabLoRA.

F.8 Partial Participation

Experimental setup. We use MNLI-m with RoBERTa-base, 50 clients, Dirichlet label skew $\alpha = 0.1$, $E = 1$ local epoch per participating client and round, and a fixed budget of 200 communication rounds. We vary the participation rate over $q \in \{1.0, 0.5, 0.2, 0.1\}$. Clients are sampled uniformly without replacement, and methods compared under the same seed use the same sampled clients.

To distinguish communication rounds from the amount of transmitted data, we also compute total parameter traffic to the 78% target. If a method communicates cP parameters per participating client and round and reaches the target in R rounds, then its total traffic is

$$RcMP = RcqNP,$$

where $M = qN$. Under the accounting in Table 17, $c = 2$ for FedProx and FedDyn and $c = 3$ for SCAFFOLD and FedStabLoRA. FedStabLoRA uses the fewest rounds at every tested participation rate. Although it communicates one additional correction vector per participating client and round, the reduction in communication rounds is sufficient to offset this extra cost in all four settings. Consequently, FedStabLoRA also achieves the lowest total parameter traffic to the 78% target across all tested participation rates.

F.9 Scaling to Larger Backbones

Experimental setup. We evaluate Qwen2.5-1.5B, 3B, and 7B on MNLI-m with 10 clients, full participation, Dirichlet label skew $\alpha = 0.1$, $E = 1$ local epoch per round, and 50 communication rounds. LoRA uses rank 8 and scaling $\alpha_{\mathrm{LoRA}} = 16$. We compare FedAvg, FedDyn, SCAFFOLD, and FedStabLoRA and tune the learning rate separately for each method and model size on *Tune*.

Table 12: MNLI-m performance under partial participation. All runs use 50 clients, Dirichlet label skew ($\alpha = 0.1$), $E = 1$, and a 200-round budget. Accuracy and rounds-to-target are mean $\pm$ standard deviation over three seeds. Total communication to 78% accuracy is derived from the same runs and reported in units of NP communicated parameters, where $N = 50$ and P is one trainable parameter vector. NR denotes that the target is not reached within 200 rounds.

Method	$q = 1.0$	$q = 0.5$	$q = 0.2$	$q = 0.1$
Accuracy (%)				
FedAvg	76.11±1.36	75.48±1.60	73.28±1.84	72.16±2.02
FedProx	82.77±0.51	82.56±0.85	81.63±1.43	81.71±1.21
FedDyn	83.07±0.85	82.04±1.50	81.31±2.32	80.74±2.02
SCAFFOLD	82.50±1.48	82.72±1.27	81.98±2.05	81.34±2.13
FedStabLoRA	86.76±0.62	85.37±0.80	84.74±1.37	84.51±1.19
Communication rounds to 78% accuracy (lower is better)				
FedAvg	NR	NR	NR	NR
FedProx	49.3±4.0	58.0±2.0	119.3±3.5	136.3±15.6
FedDyn	36.0±3.6	57.7±7.8	84.0±5.3	140.7±11.6
SCAFFOLD	39.7±5.0	51.7±6.5	81.3±5.9	123.0±8.9
FedStabLoRA	**17.3**±2.5	**37.3**±4.9	**48.3**±5.7	**84.0**±10.6
Total communication to 78% accuracy (NP parameters; lower is better)				
FedAvg	NR	NR	NR	NR
FedProx	98.6±8.0	58.0±2.0	47.7±1.4	27.3±3.1
FedDyn	72.0±7.2	57.7±7.8	33.6±2.1	28.1±2.3
SCAFFOLD	119.1±15.0	77.6±9.8	48.8±3.5	36.9±2.7
FedStabLoRA	**51.9**±7.5	**56.0**±7.4	**29.0**±3.4	**25.2**±3.2

Table 13: MNLI-m accuracy (%) across Qwen2.5 model sizes under Dirichlet label skew ($\alpha = 0.1$). All runs use 10 clients, full participation, $E = 1$, and 50 communication rounds. Results are mean $\pm$ standard deviation over three seeds.

Method	Qwen2.5-1.5B	Qwen2.5-3B	Qwen2.5-7B
FedAvg	80.34±1.22	81.35±0.96	84.31±0.06
FedDyn	82.05±0.81	82.74±0.60	85.78±0.02
SCAFFOLD	82.91±0.94	83.34±0.42	86.04±0.05
FedStabLoRA	85.28±0.73	86.87±0.83	86.34±0.08

F.10 Sensitivity Analysis

Algorithm coefficient. We use MNLI-m with RoBERTa-base, 10 clients, full participation, Dirichlet label skew $\alpha = 0.1$, $E = 2$, 50 communication rounds, and LoRA rank $r = 4$. We vary the FedStabLoRA coefficient μ while computing s from the corresponding (K, η, μ) using (4).

LoRA rank. We use the same MNLI-m setting and vary the LoRA rank over $r \in \{2, 4, 8, 16\}$. We set $\alpha_{\mathrm{LoRA}} = 2r$ so that α_{LoRA}/r remains constant. FedStabLoRA keeps the coefficient selected at $r = 4$ fixed across ranks.

F.11 Communication, Memory, and Runtime Details

Profiling setup. We profile the MNLI-m main setting with RoBERTa-base, 10 clients, full participation, Dirichlet label skew $\alpha = 0.1$, and $E = 2$. We use constant-step SGD with batch size 128 and the same mixed-precision protocol as in the main experiments: forward and backward computation use bfloat16 autocast, while trainable master parameters, federated parameter vectors, and server aggregation remain in fp32. Let $P = 740{,}355$ denote one trainable parameter vector, corresponding to 2.82 MiB (2.96 MB) in fp32. Persistent-state costs follow the implemented method states.

Runtime and peak GPU memory are measured on H200 GPUs. Each profiling job runs for five communication rounds; we discard the first round as warmup and report the median over rounds

Table 14: MNLI-m accuracy (%) for different values of the FedStabLoRA coefficient μ. Results are mean $\pm$ standard deviation over three seeds under the main MNLI-m setting.

μ	0.3	0.1	0.03	0.01
Accuracy (%)	84.30±0.15	84.23±0.77	85.07±0.65	84.32±0.74

Table 15: MNLI-m accuracy (%) for different values of the learning rate η. Results are mean $\pm$ standard deviation over three seeds under the main MNLI-m setting.

learning rate η	0.3	0.1	0.03	0.01	0.003
SCAFFOLD	DIVERGE	75.70±4.83	80.75±1.91	81.25±1.64	81.14±1.65
FedStabLoRA	85.07±0.65	84.29±0.91	82.50±0.53	82.31±0.04	81.71±0.09

2–5. Methods are launched in synchronized batches and each runtime is normalized by repeated FedAvg-LoRA runs in the same batch. The repeated FedAvg controls vary by at most 0.5% within a batch. In our simulator, one communication round executes the 10 clients sequentially on a two-GPU slot, so the normalized runtime measures relative local-computation cost rather than wall-clock time under a fully parallel federated deployment.

FedEx-LoRA residual communication. FedEx-LoRA transmits the aggregation residual from the server to the participating clients in addition to the aggregated trainable parameters. For a LoRA-adapted matrix of size $d_{\text{out}} \times d_{\text{in}}$ with rank r, the exact residual has rank at most Mr, where M is the number of participating clients (Singhal et al., 2025). It can therefore be sent in factored form instead of as a dense matrix. Let

$$P_{\text{LoRA}} := \sum_{\ell} r\big(d_{\text{out},\ell} + d_{\text{in},\ell}\big)$$

denote the number of LoRA-factor parameters across all adapted matrices, excluding the task head, and let P_{res} denote the number of parameters used to transmit the factored residual. The exact protocol gives

$$P_{\text{res}} \le M P_{\text{LoRA}}.$$

For the RoBERTa-base configuration used here, LoRA is applied to the query and value projections in all 12 layers with $r = 4$, giving $P_{\text{LoRA}} = 147{,}456$. With $M = 10$,

$$P_{\text{res}} \le 1{,}474{,}560 \approx 1.99P.$$

Thus the FedEx-LoRA downstream traffic is $P + P_{\text{res}}$, at most approximately $2.99P$ in this setting, while its upload remains P. The actual residual rank can be smaller than Mr, so we retain P_{res} symbolically in Table 17.

For FedEx-LoRA, the communication columns use protocol-level accounting. The runtime and peak-memory numbers measure our simulator implementation and do not include network-transfer time; the residual transmission is therefore counted separately through P_{res} rather than inferred from the simulator's physical message operations.

FedStabLoRA and SCAFFOLD Option II have the same state and communication footprint: each stores one additional P-dimensional client state and one server state, while each participating client downloads $2P$ and uploads $1P$. Thus, both communicate $1.5\times$ as many parameters per round as FedAvg-LoRA. FedDyn has the same persistent state size but does not require the additional downloaded correction vector.

The measured per-round runtimes of all methods remain within 1.3% of FedAvg-LoRA. FedStabLoRA is $1.013\times$ FedAvg-LoRA and SCAFFOLD Option II is $1.011\times$, consistent with both methods avoiding an additional gradient evaluation. The three stateful methods—SCAFFOLD, FedDyn, and FedStabLoRA—increase measured peak GPU memory by 36 MB over FedAvg-LoRA in the 10-client simulator, corresponding to 0.42% of the FedAvg-LoRA peak. This memory overhead scales with the number of client states stored centrally in the simulator; in a deployed federated system, each client stores only its own additional $1P$ state and the server stores one $1P$ aggregate state.

Table 16: MNLI-m accuracy (%) across LoRA ranks under Dirichlet label skew ($\alpha = 0.1$). Results are mean $\pm$ standard deviation over three seeds. The $r = 4$ column is the main MNLI-m setting.

Method	$r = 2$	$r = 4$	$r = 8$	$r = 16$
FedAvg	75.39±1.81	76.27±2.36	77.42±0.27	77.79±0.99
FedStabLoRA	83.69±0.98	**85.07±0.65**	86.62±0.41	86.42±0.17

Table 17: Per-round state, communication, runtime, and peak GPU memory on MNLI-m. State and communication are reported relative to one trainable parameter vector $P = 740{,}355$; download and upload are per participating client. For FedEx-LoRA, P_{res} denotes the additional server-to-client traffic required to transmit the aggregation residual in low-rank form. Runtime is normalized to FedAvg-LoRA within the same profiling batch. All runtime and memory measurements use the constant-SGD protocol on H200 GPUs.

Method	Client state	Server state	Download	Upload	Time/round	Peak memory
FedAvg-LoRA	0	0	$1P$	$1P$	1.000×	8622 MB
FedProx	0	0	$1P$	$1P$	1.005×	8627 MB
FedEx-LoRA	0	0	$P + P_{\text{res}}$	P	1.008×	8622 MB
FFA-LoRA	0	0	$0.90P$	$0.90P$	0.988×	7465 MB
SCAFFOLD (Option II)	$1P$	$1P$	$2P$	$1P$	1.011×	8658 MB
FedDyn	$1P$	$1P$	$1P$	$1P$	1.005×	8658 MB
FedStabLoRA	$1P$	$1P$	$2P$	$1P$	1.013×	8658 MB

G Connection to SCAFFOLD Option II

Under full participation, the $\mu = 0$ specialization of FedStabLoRA coincides with SCAFFOLD Option II (Karimireddy et al., 2020) after expressing the control variates in centered form. In this case, $s = K\eta$, and the local update becomes

$$V_{i,k+1}^t = V_{i,k}^t - \eta\big[\nabla J_i(V_{i,k}^t) - C_i^t\big].$$

The centered correction obeys the same recursion as the centered SCAFFOLD control variate, so the two methods generate the same server iterates when initialized identically.

This special case highlights an important part of our theoretical contribution. Existing analyses of SCAFFOLD assume smoothness with respect to the variables being optimized. Under LoRA, however, the optimized variables are the low-rank factors, and the smoothness of the factorized objective can grow with the factor norms even when the loss is smooth in weight space. Our analysis proves that SCAFFOLD-style correction still achieves $O(\epsilon^{-4})$ communication complexity in this setting without assuming bounded LoRA factors. Combined with the $\Theta(\epsilon^{-6})$ worst-case complexity of multi-step FedAvg(LoRA), this shows that persistent correction is sufficient to recover the centralized-order $O(\epsilon^{-4})$ rate despite factor-amplified client drift.

Our result is not limited to this SCAFFOLD-equivalent special case. For any $\mu \geq 0$, FedStabLoRA uses the penalized local update

$$V_{i,k+1}^t = V_{i,k}^t - \eta\big[\nabla J_i(V_{i,k}^t) - C_i^t + \mu(V_{i,k}^t - V^t)\big]$$

together with

$$s = \eta \sum_{k=0}^{K-1} (1 - \eta\mu)^k.$$

For $\mu > 0$, this defines a different method from SCAFFOLD Option II: the local trajectory is regularized toward the server iterate, and the correction scale is modified to match the resulting local dynamics. Theorem 5.1 establishes the same $O(\epsilon^{-4})$ communication complexity for this entire family, including both the SCAFFOLD-equivalent case $\mu = 0$ and the penalized case $\mu > 0$, without a bounded-factor assumption.

The positive penalty is therefore not needed for the asymptotic rate, but it provides an additional stabilization mechanism that is useful in practice. Our ablations show that it enlarges the stable learning-rate regime of corrected LoRA updates and improves finite-round convergence; see

Section 6.3. We further prove that the $O(\epsilon^{-4})$ rate extends to partial participation, where the Fed-StabLoRA state dynamics are analyzed directly rather than through the full-participation SCAFFOLD equivalence.

H Interpretation of Factor-Space Stationarity

Our convergence results use first-order stationarity of the factorized LoRA objective

$$J(B, A) = F(BA).$$

Writing $G = \nabla F(BA)$, the chain rule gives

$$\nabla_B J(B, A) = GA^\top, \qquad \nabla_A J(B, A) = B^\top G.$$

Thus, for a factor perturbation $(\Delta B, \Delta A)$, the corresponding first-order change in the weight update is

$$\Delta X = \Delta B\, A + B\, \Delta A,$$

and

$$\langle \nabla J(B, A), (\Delta B, \Delta A) \rangle = \langle G, \Delta B\, A + B\, \Delta A \rangle .$$

Therefore, $\nabla J(B, A) = 0$ means that there is no first-order descent direction generated by an infinitesimal perturbation of the LoRA factors. This is the stationarity notion naturally associated with gradient-based optimization of A and B, and it is the criterion used in recent convergence analyses of LoRA gradient descent.

This notion is related to, but distinct from, stationarity in weight space. When A and B have full factor rank, the directions $\Delta B\, A + B\, \Delta A$ span the tangent space of the rank-r matrix manifold at $X = BA$. In this regular case, factor-space stationarity means that $\nabla F(X)$ is orthogonal to all first-order rank-r feasible directions, i.e., X is first-order stationary for optimization restricted to rank-r matrices.

We do not claim that factor-space stationarity is equivalent to unconstrained weight-space stationarity $\nabla F(X) = 0$. In particular, when $r < \min\{m, n\}$, the weight-space gradient can contain components orthogonal to all rank-r tangent directions. Moreover, the factorization is non-identifiable: for any invertible matrix Q,

$$BA = (BQ)(Q^{-1}A),$$

so the Euclidean factor-gradient norm depends on the chosen factor representation. Rank-deficient factorizations can introduce additional degeneracies as well.

Accordingly, our theorems establish convergence to an ϵ-stationary point of the factorized objective,

$$\|\nabla J(B, A)\|_F \le \epsilon,$$

which directly characterizes the variables updated by FedAvg(LoRA) and FedStabLoRA. Obtaining a guarantee in terms of the unconstrained weight-space gradient would require additional assumptions relating the conditioning of the factors to the product BA, and is outside the scope of this work.