FEDERATED LORA UNDER CLIENT HETEROGENEITY: FACTOR REPRESENTATION AND CONVERGENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-rank adaptation (LoRA) offers a parameter-efficient approach to fine-tuning foundation models. In federated learning, LoRA can reduce local training and communication costs by restricting updates to the low-rank factors. However, the role of the low-rank parameterization in heterogeneous local training and convergence is not yet fully understood. In this paper, we develop a non-asymptotic analysis of full-batch federated LoRA with full client participation, simultaneous updates of both factors, and factor averaging at the server. We relate weight-space client heterogeneity to local drift in the LoRA factors, deriving a one-round descent bound that tracks factor norms, local steps, and stepsize. From this bound, we establish convergence rates for the minimum squared factor-gradient norm over T communication rounds. With harmonic stepsizes, we obtain an $O(1/\log T)$ rate by controlling trajectory growth, without assuming uniformly bounded factor norms. Under additional uniform trajectory bounds, a stepsize chosen according to T and held constant within each run yields an $O(T^{-2/3})$ rate. Experiments illustrate how client heterogeneity and learning-rate selection affect training dynamics, suggesting that factor and gradient magnitudes can inform local stepsize selection.

1 INTRODUCTION

Federated learning (FL) enables collaborative model training while keeping raw training data on local clients (McMahan et al., 2017). In each communication round, clients train locally and send model updates to a central server for aggregation. Although FedAvg and related methods have been widely studied (McMahan et al., 2017; Karimireddy et al., 2020b;a; Li et al., 2020), fine-tuning large foundation models across clients remains computationally and communication intensive. Advances in parameter-efficient fine-tuning of large pretrained models offer a way to address these costs in federated settings. In particular, low-rank adaptation (LoRA) (Hu et al., 2021) represents weight updates through two smaller factors, $\boldsymbol{A}$ and $\boldsymbol{B}$, while keeping the pretrained weights frozen. The resulting reduction in trainable parameters makes LoRA particularly attractive for federated fine-tuning. Integrating LoRA with FedAvg allows clients to train and exchange the LoRA factors, reducing the optimization and communication costs associated with updating the full model.

Despite these efficiency benefits, data heterogeneity remains a central challenge in federated fine-tuning. Differences in client data distributions lead to different local objectives and potentially conflicting update directions. Multiple local updates between communication rounds can accumulate client drift and slow convergence (Karimireddy et al., 2020b; Li et al., 2020). Empirical studies highlight related challenges in federated LoRA. SLoRA (Babakniya et al., 2023) reports a widening performance gap between parameter-efficient and full fine-tuning as client data heterogeneity increases. HetLoRA (Cho et al., 2024) further observes a rank-dependent trade-off between rapid performance improvement and overfitting. These findings motivate examining how the low-rank parameterization interacts with heterogeneous local training.

Recent studies have addressed aggregation discrepancies and factor alignment in federated LoRA (Wang et al., 2024c; Singhal et al., 2025; Zhang et al., 2026; Bian et al., 2025). Convergence analyses also account for heterogeneous client objectives. For example, FSLoRA assumes gradient dissimilarity directly in factor space (Fang et al., 2025), while FedRot-LoRA isolates aggregation error in its stationarity bound (Zhang et al., 2026). However, these formulations do not

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.

explicitly trace the connection from weight-space gradient heterogeneity to factor-dependent local drift and admissible stepsizes under standard factor averaging. This connection requires care because weight-space smoothness does not imply globally Lipschitz gradients in the LoRA factors. Moreover, equivalent factorizations can exhibit different factor-space gradient heterogeneity despite identical model products and weight-space client gradients. We therefore ask: *How does the LoRA factor representation shape the effect of client heterogeneity on local training and convergence?*

We study this question using full-batch local gradient descent with full client participation, where both LoRA factors are updated simultaneously and averaged separately across clients. This setting allows us to examine heterogeneous local training without stochastic-gradient or client-sampling variability. Starting from weight-space smoothness and gradient heterogeneity conditions, our analysis connects the factor representation to local drift, sufficient stepsize conditions, and convergence. Our main contributions are:

- We relate factor-space client gradient heterogeneity to weight-space gradient differences and the LoRA factor norms. We establish local-stability and drift bounds and derive a one-round descent inequality. The bounds explicitly track the interaction between client heterogeneity, factor norms, and stepsize, yielding sufficient conditions for controlling heterogeneous local updates.
- For lower-bounded nonconvex objectives, we establish non-asymptotic bounds on the minimum squared factor-gradient norm. With harmonic stepsizes, we obtain an $O(1/\log T)$ rate without assuming bounded factor norms or client gradients; instead, we derive a logarithmic growth bound on the factor norm along the trajectory. With a constant stepsize chosen according to T, we obtain $O(T^{-2/3})$ under uniform trajectory bounds.
- Guided by our convergence bounds and stepsize conditions, we design experiments to examine how learning rates, client heterogeneity, and LoRA factor scaling jointly affect training stability and convergence. The experimental results support these theoretical insights and provide practical guidance on learning-rate selection and LoRA factor initialization.

2 Related Work

Low-rank adaptation and optimization. LoRA (Hu et al., 2021) reduces the number of trainable parameters by representing weight updates as products of low-rank factors. Subsequent work improves the allocation and use of this parameter budget through adaptive rank selection, weight decomposition, and quantization (Zhang et al., 2023; Liu et al., 2024; Dettmers et al., 2023). Other developments focus on initialization (Meng et al., 2024; Wang et al., 2024a), and transformation-invariant preconditioning (Yen et al., 2024). These studies establish that LoRA optimization depends on how the low-rank update is represented and trained.

Federated optimization under heterogeneity and federated low-rank adaptation. FedAvg McMahan et al. (2017) combines multiple local optimization steps with periodic model averaging. Its analysis, together with broader results on local and decentralized optimization, characterizes how client heterogeneity and communication intervals affect convergence (Li et al., 2019; Koloskova et al., 2020). Proximal and dynamic regularization, control variates, and normalized aggregation provide complementary ways to address heterogeneous local optimization (Li et al., 2020; Acar et al., 2021; Karimireddy et al., 2020b; Wang et al., 2020). Federated instruction tuning and parameter-efficient adaptation extend language-model fine-tuning to distributed client data (Zhang et al., 2024; Babakniya et al., 2023; Wang et al., 2024c). Within this setting, a substantial line of work accommodates heterogeneous client resources through different adapter ranks or selective updates of low-rank submatrices (Cho et al., 2024; Bai et al., 2024; Fang et al., 2025).

Convergence analysis of LoRA. Existing theory studies LoRA's optimization landscape in the neural tangent kernel regime (Jang et al., 2024), establishes convergence for modified low-rank algorithms such as RAC-LoRA (Malinovsky et al., 2024), and analyzes simultaneous gradient descent on the original centralized LoRA objective (Mu & Klabjan, 2025). In distributed settings, Chen et al. (2025) formulate aggregation–broadcast operators and derive convergence conditions under bounded-factor and gradient assumptions. Convergence guarantees are also established for federated methods based on sketching and alternating optimization (Fang et al., 2025; Chen et al., 2026).

3 THEORETICAL ANALYSIS

In this section, we study the full-batch federated LoRA with full client participation. We first introduce the notation and assumptions, then derive the LoRA descent bound, control the effect of local client steps, and establish the convergence rate. All matrix norms are Frobenius norms unless otherwise specified.

3.1 PROBLEM FORMULATION AND ASSUMPTIONS

We begin with characterize the federated fine-tuning process as the model-agnostic minimization problem over N total clients:

$$\min_{\boldsymbol{W}\in\mathbb{R}^{d_1\times d_2}} f(\boldsymbol{W}_0+\boldsymbol{W}) = \frac{1}{N}\sum_{i=1}^{N} f_i(\boldsymbol{W}_0+\boldsymbol{W}), \tag{1}$$

where $\boldsymbol{W}_0 \in \mathbb{R}^{d_1\times d_2}$ represents the a frozen pretrained weight matrix, and $\boldsymbol{W}$ is the fine-tuning update. During fine-tuning, only $\boldsymbol{W}$ is optimized while $\boldsymbol{W}_0$ remains fixed. To simplify, we instead work with the following problem setup:

$$\min_{\boldsymbol{W}\in\mathbb{R}^{d_1\times d_2}} F(\boldsymbol{W}) = \frac{1}{N}\sum_{i=1}^{N} F_i(\boldsymbol{W}). \tag{2}$$

where F_i be the loss on local client i. The gradient of F_i takes the matrix form of $\nabla F_i(\boldsymbol{W})$. In the LoRA algorithm, since we parametrize $\boldsymbol{W} = \boldsymbol{B}\boldsymbol{A}$, where $\boldsymbol{B} \in \mathbb{R}^{d_1\times r}, \boldsymbol{A} \in \mathbb{R}^{r\times d_2}$ are low rank adapter matrices with rank $r < \min\{d_1, d_2\}$. This yields the factorized optimization problem

$$\min_{\boldsymbol{B}\in\mathbb{R}^{d_1\times r},\boldsymbol{A}\in\mathbb{R}^{r\times d_2}} F(\boldsymbol{B}\boldsymbol{A}) = \frac{1}{N}\sum_{i=1}^{N} F_i(\boldsymbol{B}\boldsymbol{A}). \tag{3}$$

We study full-batch local gradient descent with simultaneous updates to both factors $\boldsymbol{A}$ and $\boldsymbol{B}$. At local step k of communication round t, client i updates

$$\boldsymbol{A}^i_{t,k+1} = \boldsymbol{A}^i_{t,k} - \eta_{l,t}\nabla_{\boldsymbol{A}} F_i(\boldsymbol{B}_{t,k}\boldsymbol{A}_{t,k}), \qquad \boldsymbol{B}^i_{t,k+1} = \boldsymbol{B}^i_{t,k} - \eta_{l,t}\nabla_{\boldsymbol{B}} F_i(\boldsymbol{B}_{t,k}\boldsymbol{A}_{t,k}), \tag{4}$$

where $\eta_{l,t}$ is the learning rate at communication round t, k is the number of local steps.

Similar to many federated optimization works Li et al. (2019); Yang et al. (2021); Wang et al. (2023; 2024b), our analysis uses the following basic assumptions on the weight-space objectives.

Assumption 3.1 (Smoothness)**.** Each local objective F_i is differentiable and has an L-Lipschitz gradient: for every $i \in [N]$ and $\boldsymbol{W}_1, \boldsymbol{W}_2 \in \mathbb{R}^{d_1\times d_2}$, $\|\nabla F_i(\boldsymbol{W}_1) - \nabla F_i(\boldsymbol{W}_2)\|_F \le L\|\boldsymbol{W}_1 - \boldsymbol{W}_2\|_F$.

Assumption 3.2 (Lower-bounded objective)**.** The loss function F is lower bounded by a constant F^* such that for all $\boldsymbol{W} \in \mathbb{R}^{d_1\times d_2}, F(\boldsymbol{W}) \ge F^*$.

3.2 RESTRUCTURE THE PARAMETERS

To analyze the two factors jointly, we stack them into $\boldsymbol{V} = [\boldsymbol{B}^\top, \boldsymbol{A}]^\top$. The LoRA update can then be expressed as $g(\boldsymbol{V}) = \boldsymbol{E}_1\boldsymbol{V}\boldsymbol{V}^\top\boldsymbol{E}_2$, where $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$ are fixed selection matrices with compatible dimensions so that $g(\boldsymbol{V}) \in \mathbb{R}^{d_1\times d_2}$. The matrix $\boldsymbol{V}$ is the trainable LoRA factor. We define the corresponding factor-space objectives as $J(\boldsymbol{V}) = F(g(\boldsymbol{V})) = \frac{1}{N}\sum_{i=1}^{N} J_i(\boldsymbol{V})$ and $J_i(\boldsymbol{V}) = F_i(g(\boldsymbol{V}))$. The lower bound on F (Assumption 3.2) also gives $J(\boldsymbol{V}) \ge F^\star$; hence J is bounded below, and we may take $J^\star = F^\star$.

Let $\boldsymbol{V}_t$ denote the server iterate at communication round t, and let $\boldsymbol{V}^i_{t,k}$ denote the local iterate of client i after k steps in that round. All clients initialize from the same server iterate, $\boldsymbol{V}^i_{t,0} = \boldsymbol{V}_t$, and perform $K \ge 1$ full-batch gradient steps: $\boldsymbol{V}^i_{t,k+1} = \boldsymbol{V}^i_{t,k} - \eta_{\ell,t}\nabla J_i(\boldsymbol{V}^i_{t,k}), \quad k = 0, \dots, K-1$. The stepsize is shared by all clients and remains constant within each round. We use η_ℓ when analyzing a fixed round and retain the subscript t when considering variation across rounds.

After local training, the server averages the final factor iterates $\boldsymbol{V}_{t+1} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{V}^i_{t,K}$. The resulting trainable weight update is $g(\boldsymbol{V}_{t+1}) = \boldsymbol{E}_1\boldsymbol{V}_{t+1}\boldsymbol{V}_{t+1}^\top\boldsymbol{E}_2$. This procedure averages the LoRA

factors separately; in general, the resulting product differs from the average of the client products, $N^{-1}\sum_{i=1}^{N} g(\boldsymbol{V}_{t,K}^{i})$.

We denote the global factor update by $\boldsymbol{\Delta}_t := \boldsymbol{V}_{t+1} - \boldsymbol{V}_t$. To describe the gradient dissimilarity between the global loss function and the local loss function, we need the assumption for bounded gradient dissimilarity as follows.

Assumption 3.3 (Bounded gradient dissimilarity)**.** For a finite $\sigma_g \geq 0$, the loss function has a global variance bound ($\boldsymbol{W}$-space heterogeneity bound): $\frac{1}{N}\sum_{i=1}^{N} \|\nabla F_i(g(\boldsymbol{V}_t)) - \nabla F(g(\boldsymbol{V}_t))\|_F^2 \leq \sigma_g^2$.

Assumption 3.3 is a common assumption in analyzing federated non-convex optimization methods Li et al. (2019); Yang et al. (2021); Wang et al. (2023; 2024b).

3.3 Main results

The analysis proceeds in three steps. We first derive a descent bound for the LoRA composition that accounts for higher-order update terms. We then relate client heterogeneity to factor-space gradients and bound the drift accumulated during local training. Finally, we derive a one-round descent bound and establish convergence rates for harmonic and constant local stepsizes. The supporting lemmas and detailed proofs are provided in Appendix A.

Step 1: LoRA composition and descent. The composition $J_i(\boldsymbol{V}) = F_i(g(\boldsymbol{V}))$ has the same form as the centralized LoRA objective studied by Mu & Klabjan (2025). Their descent lemma therefore applies to each client objective and, since F is also L-smooth, to the global objective.

Lemma 3.4 (LoRA descent lemma for a client)**.** *Suppose F_i is L-smooth. Then for any $\boldsymbol{V}_1, \boldsymbol{V}_2 \in \mathbb{R}^{(d_1+d_2)\times r}$,*

$$J_i(\boldsymbol{V}_2) \leq J_i(\boldsymbol{V}_1) + \langle \nabla J_i(\boldsymbol{V}_1), \boldsymbol{V}_2 - \boldsymbol{V}_1 \rangle_F + \sqrt{2}L\|\boldsymbol{V}_2 - \boldsymbol{V}_1\|_F^2\|\boldsymbol{V}_1\|_F^2 + \sqrt{2}L\|\boldsymbol{V}_2 - \boldsymbol{V}_1\|_F^3\|\boldsymbol{V}_1\|_F + \frac{\sqrt{2}L}{4}\|\boldsymbol{V}_2 - \boldsymbol{V}_1\|_F^4 + \|\nabla F_i(g(\boldsymbol{V}_1))\|_F\|\boldsymbol{V}_2 - \boldsymbol{V}_1\|_F^2. \tag{5}$$

Step 2: Heterogeneity and local drift. A communication round combines gradients evaluated at different local iterates. We first quantify how the LoRA parameterization relates weight-space gradient heterogeneity to gradient heterogeneity in factor space.

Lemma 3.5 (Heterogeneity amplification)**.** *For any $\boldsymbol{V}$ satisfying the weight-space heterogeneity bound in Assumption 3.3, in particular any iterate $\boldsymbol{V}_t$,*

$$\frac{1}{N}\sum_{i=1}^{N} \|\nabla J_i(\boldsymbol{V}) - \nabla J(\boldsymbol{V})\|_F^2 \leq 4\|\boldsymbol{V}\|_F^2\sigma_g^2. \tag{6}$$

Lemma 3.5 shows that the factor-space heterogeneity bound depends on both weight-space heterogeneity and the LoRA factor norm. This dependence reflects the role of the factor representation: equivalent factorizations of the same model can exhibit different factor-space heterogeneity, as illustrated by scalar rescaling in Section 3.4. We next use the bound to control the drift accumulated during local training.

Lemma 3.6 (Average local drift)**.** *With Lemma A.3 holds and $6K\eta_\ell^2\Lambda_t \leq \frac{1}{2K-1}$ holds, where $G_t^2 := \max_{i\in[N]} \|\nabla F_i(g(\boldsymbol{V}_t))\|_F^2$, and $\Lambda_t := 288L^2\|\boldsymbol{V}_t\|_F^4 + 8G_t^2$. Then for every $k = 0, \ldots, K$,*

$$\frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{V}_{t,k}^{i} - \boldsymbol{V}_t\|_F^2 \leq 30K^2\eta_\ell^2\left(\|\nabla J(\boldsymbol{V}_t)\|_F^2 + 4\|\boldsymbol{V}_t\|_F^2\sigma_g^2\right). \tag{7}$$

Step 3: One-round descent and convergence. Combining the drift and update-norm bounds with the global descent template in Lemma A.1 yields the following one-round bound.

Lemma 3.7 (One-round descent)**.** *Under Assumptions 3.1 and 3.3, let $h_t = \eta_{\ell,t}K$, if $h_t^2\Lambda_t \leq 1/60$, then*

$$J(\boldsymbol{V}_{t+1}) \leq J(\boldsymbol{V}_t) - \frac{h_t}{4}\|\nabla J(\boldsymbol{V}_t)\|_F^2 - \frac{\|\boldsymbol{\Delta}_t\|_F^2}{4h_t} + 60h_t^3\Lambda_t\|\boldsymbol{V}_t\|_F^2\sigma_g^2. \tag{8}$$

The negative update-norm term controls trajectory growth, while the cubic heterogeneity residual enables the convergence bounds below.

Theorem 3.8 (Convergence with harmonic local stepsizes)**.** *Suppose Assumptions 3.1, 3.2, and 3.3 hold. Consider T rounds of full-batch federated LoRA with full participation, factor averaging, and a fixed number $K \geq 1$ of local steps. Choose $\eta_{\ell,t} = \frac{\gamma_0}{K(t+1)(1+\sigma_g)}, t = 0, \ldots, T-1$, where $\gamma_0 > 0$ is a sufficiently small constant independent of T, as specified in Appendix A.4.2. With L, γ_0, and the initial objective gap held fixed,*

$$\min_{0 \leq t < T} \|\nabla J(\boldsymbol{V}_t)\|_F^2 = O\left(\frac{1+\sigma_g}{\log(T+1)}\right). \tag{9}$$

Theorem 3.8 establishes an $O(1/\log T)$ stationarity rate. This has the same order in T as the general centralized gradient-descent rate, i.e., we obtain the same $O(1/\log T)$ rate as Mu & Klabjan (2025) without assuming bounded factors, but the mechanism differs. In the federated analysis, Lemma 3.5 relates weight-space heterogeneity to factor-space heterogeneity through the bound $4\|\boldsymbol{V}\|_F^2\sigma_g^2$. Under the stated conditions, the harmonic schedule makes the accumulated residual finite, so heterogeneity affects the numerator of the stationarity bound without introducing a non-vanishing residual term.

High-level stepsize requirements. The local stepsize must ensure stability and control the drift accumulated over K local updates. A sufficient local-stability condition has the form

$$\eta_{\ell,t} K \lesssim \frac{1}{L\|\boldsymbol{V}_t\|_F^2 + G_t}, \quad \text{where } G_t^2 := \max_{i \in [N]} \|\nabla F_i(g(\boldsymbol{V}_t))\|_F^2, \tag{10}$$

where the implicit constant is sufficiently small and universal. Larger factor norms or maximum client-gradient norms therefore tighten this sufficient condition. The factor-space gradient and client heterogeneity enter the drift bound through $\|\nabla J(\boldsymbol{V}_t)\|_F^2 + 4\|\boldsymbol{V}_t\|_F^2\sigma_g^2$.

The logarithmic growth bound relies on the decaying harmonic schedule. For a constant stepsize $\eta_\ell K = \gamma T^{-1/3}$, the same argument only gives $\|\boldsymbol{V}_t\|_F^2 = O(T^{2/3})$, so $\eta_\ell^2 K^2 \Lambda_t$ may grow with T and the stepsize condition can no longer be verified from the derived bound. For this schedule, we therefore impose the following assumption.

Assumption 3.9 (Uniform trajectory bounds)**.** For the stepsize schedules considered below, there exist finite constants $B_V, B_G \geq 0$, independent of the total number of communication rounds T, such that the generated server iterates satisfy

$$\|\boldsymbol{V}_t\|_F^2 \leq B_V, \qquad G_t^2 = \max_{i \in [N]} \|\nabla F_i(g(\boldsymbol{V}_t))\|_F^2 \leq B_G^2, \qquad 0 \leq t < T. \tag{11}$$

Bounded-factor conditions are also used in existing analyses of federated and decentralized low-rank fine-tuning (Guo et al., 2025; Ghiasvand et al., 2025). The following result complements Theorem 3.8 by choosing a stepsize that is constant across communication rounds.

Theorem 3.10 (Convergence with a constant local stepsize)**.** *Suppose Assumptions 3.1, 3.2, 3.3 and 3.9 hold. Consider the full-batch federated LoRA with full participation and factor averaging. Choose $\eta_\ell = \frac{\gamma}{KT^{1/3}}$, where $\gamma > 0$ is a sufficiently small constant independent of T. With L, B_V, B_G, γ, and the initial objective gap held fixed,*

$$\min_{0 \leq t < T} \|\nabla J(\boldsymbol{V}_t)\|_F^2 = O\left(\frac{1+\sigma_g^2}{T^{2/3}}\right). \tag{12}$$

The stepsize $\gamma/(KT^{1/3})$ depends on the prescribed number of rounds T and the number of local steps K, but remains fixed within each run. A fixed positive stepsize independent of T instead leaves an $O((\eta_\ell K)^2\sigma_g^2)$ residual term in the bound.

3.4 Additional Discussion

Stochastic gradients and partial participation. We focus on full-batch local gradient descent with full participation to separate the effect of data heterogeneity from stochastic-gradient and client-sampling noise. Both extensions face a common obstacle. Our harmonic analysis derives the factor-norm growth bound from a one-round descent inequality that holds deterministically in every round. With stochastic gradients or client sampling, this inequality holds only in expectation, and the noise itself grows with the factor norm. For example, under uniform client sampling, the sampling variance grows with the dispersion of client updates, which scales with the factor-space heterogeneity bound $4\|\boldsymbol{V}_t\|_F^2\sigma_g^2$ in Lemma 3.5. Extending the analysis would therefore require controlling the factor-norm growth in expectation or with high probability, for example through a stopping-time argument, as well as the cubic and quartic update terms in $\|\boldsymbol{\Delta}_t\|_F$. We leave these extensions to future work.

Scalar rescaling and convergence. Scalar rescaling illustrates how factor-space gradient heterogeneity depends on the LoRA representation, even when the model product is fixed. Following Definition 2 of Yen et al. (2024), consider a nonzero scalar s and the equivalent factors$\boldsymbol{B}_s = s\boldsymbol{B}, \quad \boldsymbol{A}_s = s^{-1}\boldsymbol{A}, \quad \boldsymbol{B}_s\boldsymbol{A}_s = \boldsymbol{B}\boldsymbol{A}$. Their stacked representation is $\boldsymbol{V}_s = [\boldsymbol{B}_s^\top, \boldsymbol{A}_s]^\top$, and $J_i(\boldsymbol{V}_s) = J_i(\boldsymbol{V})$ holds for every client. By applying the chain rule

$$\begin{aligned} \nabla_{\boldsymbol{B}} J_i(\boldsymbol{V}) &= \nabla F_i(\boldsymbol{B}\boldsymbol{A})\boldsymbol{A}^\top, & \nabla_{\boldsymbol{A}} J_i(\boldsymbol{V}) &= \boldsymbol{B}^\top \nabla F_i(\boldsymbol{B}\boldsymbol{A}) \\ \nabla_{\boldsymbol{B}_s} J_i(\boldsymbol{V}_s) &= s^{-1}\nabla_{\boldsymbol{B}} J_i(\boldsymbol{V}), & \nabla_{\boldsymbol{A}_s} J_i(\boldsymbol{V}_s) &= s\nabla_{\boldsymbol{A}} J_i(\boldsymbol{V}). \end{aligned} \tag{13}$$

With the same local stepsize, equivalent factorizations can thus produce different factor updates and optimization trajectories.

The rescaled factor norm satisfies $\|\boldsymbol{V}_s\|_F^2 = s^2\|\boldsymbol{B}\|_F^2 + s^{-2}\|\boldsymbol{A}\|_F^2$. The corresponding factor-space gradient heterogeneity satisfies the exact identity

$$\begin{aligned} &\frac{1}{N}\sum_i \|\nabla J_i(\boldsymbol{V}_s) - \nabla J(\boldsymbol{V}_s)\|_F^2 \\ &= \frac{1}{Ns^2}\sum_i \|\nabla_{\boldsymbol{B}} J_i(\boldsymbol{V}) - \nabla_{\boldsymbol{B}} J(\boldsymbol{V})\|_F^2 + \frac{s^2}{N}\sum_i \|\nabla_{\boldsymbol{A}} J_i(\boldsymbol{V}) - \nabla_{\boldsymbol{A}} J(\boldsymbol{V})\|_F^2. \end{aligned} \tag{14}$$

Thus, a common scalar rescaling reweights client gradient differences in the two factor blocks. Equivalent factorizations can consequently exhibit different factor-space heterogeneity despite identical weight-space client gradients.

This dependence on the factor representation also enters the convergence analysis. At a fixed model product, rescaling leaves G_t unchanged but can change $\|\boldsymbol{V}_t\|_F$. The factor norm enters both $\Lambda_t = 288L^2\|\boldsymbol{V}_t\|_F^4 + 8G_t^2$ and the heterogeneity upper bound $4\|\boldsymbol{V}_t\|_F^2\sigma_g^2$, thereby affecting the local-stability and drift bounds and the sufficient stepsize conditions used in Theorems. When the factor-norm term dominates Λ_t, $1/\sqrt{\Lambda_t}$ scales approximately as $1/\|\boldsymbol{V}_t\|_F^2$ for fixed L. Larger factor norms therefore tighten this part of the sufficient stepsize condition.

4 Experiments

4.1 Experimental Settings

General settings. We evaluate federated LoRA in three settings: using an ImageNet-pretrained ResNet-18 (He et al., 2016) as a frozen feature extractor and train a LoRA-parameterized multinomial logistic regression classifier on CIFAR-10 (Krizhevsky et al., 2009), LoRA fine-tuning of a pretrained ResNet-18 on CIFAR-10, and Qwen2.5-1.5B-Instruct model (Team, 2024) fine-tuning on commonsense reasoning. Both ResNet-18 backbones are pretrained on ImageNet. The CIFAR-10 experiments use 100 clients with i.i.d. or class-wise Dirichlet partitions, with $\alpha_{\mathrm{Dir}} \in \{0.1, 0.5\}$. The language-model benchmark combines ARC-Easy Clark et al. (2018), ARC-Challenge Clark et al. (2018), BoolQ Clark et al. (2019), HellaSwag Zellers et al. (2019), and PIQA Bisk et al. (2020), using 5,000 training and 1,000 held-out examples, with 10 total clients, and task-level heterogeneity controlled by $\alpha_{\mathrm{Dir}} = 0.1$.

Federated training. Clients initialize from the current global model, update both LoRA factors, and return them for separate factor averaging. We use sample-weighted aggregation for CIFAR-10 and uniform aggregation for the equal-size language-model clients, with server learning rate 1. Both

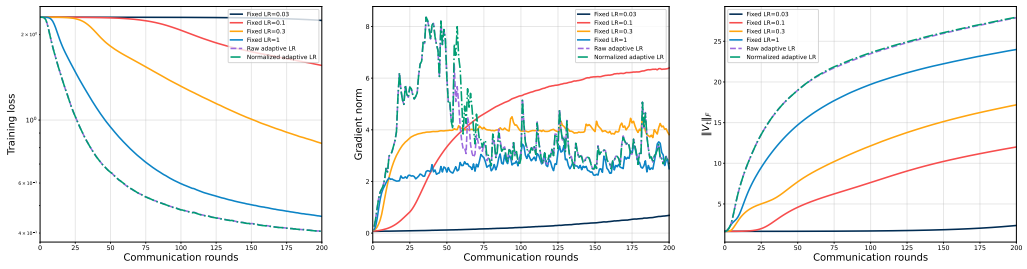
CIFAR-10 experiments use rank-8 LoRA and $K = 10$ local SGD steps with batch size 32. We run 200 rounds with full participation or 500 rounds with 10% participation. Qwen2.5 uses rank-16 LoRA on query and value projections, with 20 rounds for full participation and 50 rounds with 20% participation. Each selected client performs 20 AdamW steps with effective batch size 4 and learning rate 2×10^{-4}.

Evaluation and sensitivity. For logistic regression, we vary the learning rate and initial factor scale while preserving the initial product $\boldsymbol{BA}$. We monitor training and validation performance and reserve the CIFAR-10 test set for final evaluation. For language-model fine-tuning, we report per-task, macro, and worst-task accuracy, together with factor norms, update norms, and factor-gradient heterogeneity. Experiments run on an NVIDIA RTX A6000 GPU. Full preprocessing, optimizer, implementation, and diagnostic details are provided in Appendix B.

Adaptive local learning-rate design For norm-adaptive learning-rate experiments, with the definition in section 3.1: $V_t = [B_t^\top, A_t]$, $\|V_t\|_F^2 = \|A_t\|_F^2 + \|B_t\|_F^2$, where V_t denotes the global LoRA factors before communication round t. For every selected client i, we evaluate the LoRA gradient on the first local mini-batch without updating the model. We then use the maximum squared gradient norm across participating clients, $\bar{g}_t = \max_{i \in \mathcal{S}_t} \|\nabla F_i(g(V_t))\|_F^2$, and smooth it using an exponential moving average, $g_t = \beta g_{t-1} + (1-\beta)\bar{g}_t$, with $\beta = 0.9$. Using the maximum provides a conservative estimate of the round-wise gradient scale. We first use the base local learning rate η_0 for the first 10% warmup rounds. Rather than using potentially noisy statistics at initialization, we define the reference quantities $\|V_{\text{ref}}\|_F^2$ and g_{ref} by averaging the factor and gradient statistics over the final five warmup rounds (and the final warmup round for LLM experiment). We consider two strategies for combining these quantities. The raw variant directly adds the squared factor and gradient norms: $m_t^{\text{raw}} = \frac{\|V_{\text{ref}}\|_F^2 + g_{\text{ref}} + \epsilon}{\|V_t\|_F^2 + g_t + \epsilon}$. This rule decreases the learning rate as the combined factor-and-gradient scale grows. However, because the two quantities are combined in their original scales, the larger component may dominate the schedule. The normalized variant first expresses each component relative to its own reference value: $m_t^{\text{norm}} = 2 \cdot \left(\frac{\|V_t\|_F^2 + \epsilon}{\|V_{\text{ref}}\|_F^2 + \epsilon} + \frac{g_t + \epsilon}{g_{\text{ref}} + \epsilon} \right)^{-1}$. For either variant, the local learning rate is $\eta_t = \eta_0 \max(m_t, m_{\min})$, where $\epsilon = 10^{-12}$. The same round-wise learning rate is used by all participating clients.

4.2 Main Results

We examine how learning-rate selection and client heterogeneity affect optimization and final test accuracy in the two CIFAR-10 settings. The learning-rate comparisons use the Dirichlet partition with $\alpha_{\text{Dir}} = 0.1$.



(a) Training loss (b) Gradient norm (c) $\boldsymbol{V}$ norm

Figure 1: Full participation of logistic regression experiment

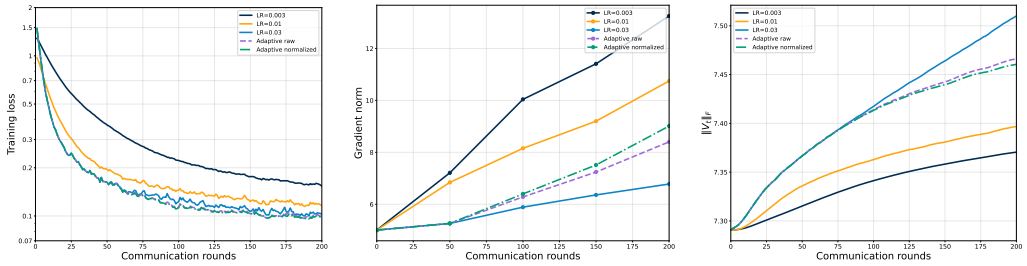
Learning-rate sensitivity in logistic regression. The logistic-regression experiments showed a clear sensitivity to the local learning rate (see Figures 1). Small fixed rates reduced the training loss slowly, whereas $\eta_\ell = 1$ and both adaptive schedules achieved substantially lower losses within the available rounds. The raw and normalized schedules retained this rapid loss reduction under both full and partial participation. As reported in Table 2, both adaptive variants improved final test

Table 1: LLM fine-tuning results after 20 communication rounds with full client participation.

Method	Final training loss	Macro accuracy (%)	Worst-task accuracy (%)
Fixed LR	0.2058	**82.0**	76.5
Adaptive raw	0.2024	81.8	76.5
Adaptive normalized	**0.2001**	**82.0**	**77.0**

accuracy over $\eta_\ell = 1$ under full participation. Their gains under partial participation were 0.83 and 0.92 percentage points, respectively. [1]

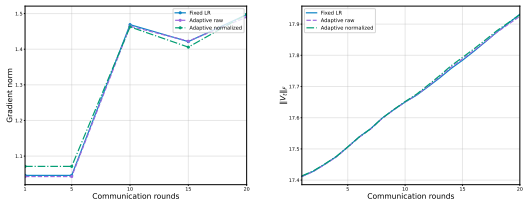
ResNet-18 fine-tuning. The ResNet-18 experiments also showed sensitivity to the fixed learning rate (Figures 2). The $\eta_\ell = 0.003$ baseline reduced training loss more slowly than $\eta_\ell = 0.03$, while the adaptive schedules achieved losses close to the larger fixed-rate baseline. However, their final test accuracies remained numerically close to the fixed-rate result, differing by at most 0.22 percentage points across the two participation regimes. The fixed rate yielded the highest mean accuracy under full participation, while the raw adaptive schedule yielded the highest mean under partial participation.



(a) Training loss (b) Gradient norm (c) $\boldsymbol{V}$ norm

Figure 2: Full participation of fine-tuning ResNet-18 experiment

LLM fine-tuning. As shown in Table 1, adaptive schedules yielded slightly lower final training losses, while macro accuracy remained numerically similar across methods. The normalized adaptive schedule matched the fixed learning rate in macro accuracy, and had slightly higher worst-task accuracy.



(a) Gradient norm (b) $\boldsymbol{V}$ norm

Figure 3: Norm diagnostics for Qwen2.5-1.5B-Instruct fine-tuning over 20 rounds with full client participation.

Effect of client heterogeneity. Logistic-regression training losses, ResNet-18 model fine-tuning losses and fine-tuning of Qwen2.5-1.5B-Instruct model were higher on more heterogeneous client partitions under both participation regimes (Figure 8 in Appendix B). This difference also appeared in final test accuracy. Logistic-regression accuracy on $\alpha_{\mathrm{Dir}} = 0.1$ data was 2.47 percentage points lower than on i.i.d. data under full participation and 5.63 points lower under partial participation (Table 2). For ResNet-18 with partial participation, the corresponding reduction was 3.86 points, with $\alpha_{\mathrm{Dir}} = 0.5$ yielding intermediate accuracy. These comparisons associate stronger client heterogeneity with lower accuracy in both adaptation settings.

Norm diagnostics and adaptive learning rates. Figures 1, Figure 2 and Figure 3 show that the factor norms $\|\boldsymbol{V}_t\|_F$ generally increased as training loss decreased. Smaller factor norms alone therefore did not identify the lowest-loss runs. The gradient panels report an estimate of the maximum

[1] Due to space limitation, we leave the training loss, gradient norm and $\boldsymbol{V}$ norm comparison of partial participation in the Appendix B.

participating-client squared weight-space gradient norm, $\bar{g}_t$. Although this quantity differs from the global factor-gradient norm in our convergence theorems, it captures changes in client-level gradient magnitudes during training.

In logistic regression, both adaptive schedules rapidly reduced training loss and improved mean final test accuracy over the fixed $\eta_\ell = 1$ baseline. Accuracy gains were less consistent for ResNet-18, while LLM fine-tuning showed norm trajectories similar to those of the fixed-rate baseline. Overall, the results suggest that factor and gradient magnitudes can inform learning-rate selection, although the benefits vary across adaptation settings.

Table 2: Comparison of different methods under full and partial settings on two image classification tasks.

Logistic Regression			**ResNet-18**		
Learning rate	Full	Partial	Learning rate	Full	Partial
Fixed $\eta_l = 1$	83.40±0.10	80.69±1.32	Fixed $\eta_l = 0.03$	**91.16**±0.28	89.60±0.59
Raw adaptive LR	**84.68**±0.10	81.52±1.00	Raw adaptive LR	90.94±0.35	**89.72**±0.78
Normalized adaptive LR	84.67±0.10	**81.61**±0.98	Normalized adaptive LR	90.99±0.40	89.56±0.74
Heterogeneity level	Full	Partial	Heterogeneity level	Full	Partial
$\alpha_{\text{Dir}} = 0.1$	83.40±0.10	80.69±1.32	$\alpha_{\text{Dir}} = 0.1$	91.16±0.28	89.60±0.59
$\alpha_{\text{Dir}} = 0.5$	85.78±0.10	85.49±0.45	$\alpha_{\text{Dir}} = 0.5$	92.09±0.20	92.39±0.14
i.i.d.	**85.87**±0.04	**86.32**±0.23	i.i.d.	**92.65**±0.08	**93.46**±0.08

Ablation on the scalar s We investigate the interaction between the learning rate and the relative scaling of the LoRA factors by applying the transformation $\boldsymbol{A} \leftarrow \boldsymbol{A}/s$ and $\boldsymbol{B} \leftarrow s\boldsymbol{B}$, which leaves the effective update BA unchanged. We vary $s \in \{0.25, 0.5, 1, 2, 4, 8, 16, 32\}$ under partial participation in the logistic regression experiment. With $\eta_\ell = 0.1$, the terminal training loss varies substantially, decreasing as s increases from 0.25 to 8, before increasing again at larger scales. In contrast, $\eta_\ell = 1$ remains relatively stable across moderate values of s. These results demonstrate that optimization is sensitive to the relative scaling of the LoRA factors, particularly at smaller learning rates, highlighting a nontrivial interaction between factor parameterization and learning-rate selection.

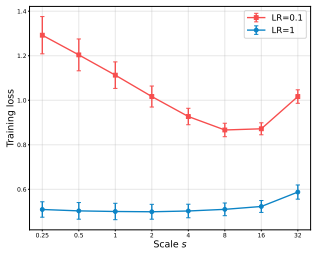


Figure 4: Ablation on scalar s.

5 Conclusion

In this paper, we analyzed the interaction between client heterogeneity and the LoRA factor representation under full-batch local gradient descent, full client participation, and factor averaging. The analysis shows how factor-space gradient heterogeneity depends on the chosen factorization, even when the represented weight update remains unchanged. By tracing this dependence through local drift and one-round descent, we establish non-asymptotic stationarity bounds and sufficient stepsize conditions. For harmonic stepsizes, the $O(1/\log T)$ bound does not require bounded factor norms, since the growth of the factor norm is derived along the trajectory. Experiments illustrate the effects of heterogeneity and learning-rate selection and provide empirical support for adapting stepsizes to factor and gradient magnitudes. Extending these guarantees to stochastic local updates and partial client participation remains an important direction for future work.

AI USE STATEMENT

We used generative AI tools primarily for coding assistance and manuscript editing to improve clarity and readability. Additional assistance included drafting selected passages, identifying relevant literature, formatting references, and checking the consistency of theoretical statements and their assumptions. The authors take full responsibility for the accuracy and integrity of the final manuscript, code, and reported results. The authors reviewed the AI-assisted text for technical accuracy and checked and tested the AI-assisted code for correctness.

REFERENCES

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas, Matthew Mattina, Paul Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=B7v4QMR6Z9w.

Sara Babakniya, Ahmed Roushdy Elkordy, Yahya H Ezzeldin, Qingfeng Liu, Kee-Bong Song, Mostafa El-Khamy, and Salman Avestimehr. Slora: Federated parameter efficient fine-tuning of language models. *arXiv preprint arXiv:2308.06522*, 2023.

Jiamu Bai, Daoyuan Chen, Bingchen Qian, Liuyi Yao, and Yaliang Li. Federated fine-tuning of large language models under heterogeneous language tasks and client resources. *arXiv e-prints*, pp. arXiv–2402, 2024.

Jieming Bian, Lei Wang, Letian Zhang, and Jie Xu. Lora-fair: Federated lora fine-tuning with aggregation and initialization refinement. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3737–3746. IEEE, 2025.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 7432–7439, 2020.

Shuangyi Chen, Yuanxin Guo, Yue Ju, Hardik Dalal, Zhongwen Zhu, and Ashish Khisti. Robust federated finetuning of llms via alternating optimization of lora. *Advances in Neural Information Processing Systems*, 38:120038–120090, 2026.

Xin Chen, Shuaijun Chen, Omid Tavallaie, Nguyen Tran, Shuhuang Xiang, and Albert Zomaya. Convergence analysis of aggregation-broadcast in lora-enabled distributed fine-tuning. *arXiv preprint arXiv:2508.01348*, 2025.

Yae Jee Cho, Luyang Liu, Zheng Xu, Aldi Fahrezi, and Gauri Joshi. Heterogeneous low-rank approximation for federated fine-tuning of on-device foundation models. *arXiv preprint arXiv:2401.06432*, 2024.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, 2019.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

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. *arXiv preprint arXiv:2501.19389*, 2025.

Sajjad Ghiasvand, Mahnoosh Alizadeh, and Ramtin Pedarsani. Decentralized low-rank fine-tuning of large language models. In *Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025)*, pp. 334–345, 2025.

Pengxin Guo, Shuang Zeng, Yanran Wang, Huijie Fan, Feifei Wang, and Liangqiong Qu. Selective aggregation for low-rank adaptation in federated learning. In *International Conference on Learning Representations*, volume 2025, pp. 99003–99027, 2025.

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

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. *arXiv preprint arXiv:2106.09685*, 2021.

Uijeong Jang, Jason D. Lee, and Ernest K. Ryu. LoRA training in the NTK regime has no spurious local minima. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 21306–21328. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/jang24d.html.

Sai Praneeth Karimireddy, Martin Jaggi, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Mime: Mimicking centralized stochastic algorithms in federated learning. *arXiv preprint arXiv:2008.03606*, 2020a.

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

Anastasia Koloskova, Nicolas Loizou, Sadra Boreiri, Martin Jaggi, and Sebastian Stich. A unified theory of decentralized sgd with changing topology and local updates. In *International Conference on Machine Learning*, pp. 5381–5393. PMLR, 2020.

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

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

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353*, 2024.

Grigory Malinovsky, Umberto Michieli, Hasan Abed Al Kader Hammoud, Taha Ceritli, Hayder Elesedy, Mete Ozay, and Peter Richtárik. Randomized asymmetric chain of lora: The first meaningful theoretical framework for low-rank adaptation. *arXiv preprint arXiv:2410.08305*, 2024.

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

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. Pissa: Principal singular values and singular vectors adaptation of large language models. *Advances in Neural Information Processing Systems*, 37:121038–121072, 2024.

Siqiao Mu and Diego Klabjan. On the convergence rate of LoRA gradient descent, 2025. URL https://arxiv.org/abs/2512.18248. Version 3, revised 11 May 2026.

Raghav Singhal, Kaustubh Ponkshe, and Praneeth Vepakomma. Fedex-lora: Exact aggregation for federated and efficient fine-tuning of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1316–1336, 2025.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *arXiv preprint arXiv:2007.07481*, 2020.

Shaowen Wang, Linxi Yu, and Jian Li. Lora-ga: Low-rank adaptation with gradient approximation. *Advances in Neural Information Processing Systems*, 37:54905–54931, 2024a.

Yujia Wang, Yuanpu Cao, Jingcheng Wu, Ruoyu Chen, and Jinghui Chen. Tackling the data heterogeneity in asynchronous federated learning with cached update calibration. In *Federated Learning and Analytics in Practice: Algorithms, Systems, Applications, and Opportunities*, 2023.

Yujia Wang, Shiqiang Wang, Songtao Lu, and Jinghui Chen. Fadas: Towards federated adaptive asynchronous optimization. *arXiv preprint arXiv:2407.18365*, 2024b.

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. *arXiv preprint arXiv:2409.05976*, 2024c.

Haibo Yang, Minghong Fang, and Jia Liu. Achieving linear speedup with partial worker participation in non-iid federated learning. *arXiv preprint arXiv:2101.11203*, 2021.

Jui-Nan Yen, Si Si, Zhao Meng, Felix Yu, Sai Surya Duvvuri, Inderjit S. Dhillon, Cho-Jui Hsieh, and Sanjiv Kumar. LoRA done RITE: Robust invariant transformation equilibration for LoRA optimization, 2024. URL https://arxiv.org/abs/2410.20625v1. Version 1.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4791–4800, 2019.

Haoran Zhang, Dongjun Kim, Seohyeon Cha, and Haris Vikalo. Fedrot-lora: Mitigating rotational misalignment in federated lora. *arXiv preprint arXiv:2602.23638*, 2026.

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

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*, 2023.

A Proofs and Additional Results for Theoretical Analysis

This appendix provides the supporting lemmas and derivations for the results in the main text. The notation and assumptions are those of the theoretical analysis.

A.1 LoRA identities and the global descent template

A.1.1 Proof of Lemma 3.4

Proof. Apply Lemma 3.3 of Mu & Klabjan (2025) with the original loss replaced by F_i. Its composition is exactly $J_i(\boldsymbol{V}) = F_i(\boldsymbol{E}_1\boldsymbol{V}\boldsymbol{V}^\top\boldsymbol{E}_2)$, and its smoothness requirement is the assumed L-smoothness of F_i. This gives the displayed inequality with the same constants. Since $F = N^{-1}\sum_i F_i$ is also L-smooth, the same argument applies to $J(\boldsymbol{V}) = F(g(\boldsymbol{V}))$. □

Lemma A.1 (Global one-round descent template)**.** *Suppose the global objective F is L-smooth. Then the averaged iterate satisfies*

$$\begin{aligned} J(\boldsymbol{V}_{t+1}) \le J(\boldsymbol{V}_t) &- \frac{\eta_\ell K}{2}\|\nabla J(\boldsymbol{V}_t)\|_F^2 + \frac{\eta_\ell}{2}\sum_{k=0}^{K-1}\|\boldsymbol{D}_{t,k}\|_F^2 \\ &+ \sqrt{2}L\|\boldsymbol{\Delta}_t\|_F^2\|\boldsymbol{V}_t\|_F^2 + \sqrt{2}L\|\boldsymbol{\Delta}_t\|_F^3\|\boldsymbol{V}_t\|_F \\ &+ \frac{\sqrt{2}L}{4}\|\boldsymbol{\Delta}_t\|_F^4 + \|\nabla F(g(\boldsymbol{V}_t))\|_F\|\boldsymbol{\Delta}_t\|_F^2. \end{aligned} \tag{15}$$

Proof. Apply Lemma 3.4 to the global objective $J(\boldsymbol{V}) = F(g(\boldsymbol{V}))$ with $\boldsymbol{V}_1 = \boldsymbol{V}_t$ and $\boldsymbol{V}_2 = \boldsymbol{V}_{t+1}$. It remains to expand the linear term. Since

$$\boldsymbol{\Delta}_t = -\eta_\ell \sum_{k=0}^{K-1}\frac{1}{N}\sum_{i=1}^{N}\nabla J_i(\boldsymbol{V}_{t,k}^i),$$

and

$$\frac{1}{N}\sum_{i=1}^{N}\nabla J_i(\boldsymbol{V}_{t,k}^i) = \nabla J(\boldsymbol{V}_t) + \boldsymbol{D}_{t,k},$$

we obtain

$$\langle \nabla J(\boldsymbol{V}_t), \boldsymbol{\Delta}_t\rangle_F = -\eta_\ell \sum_{k=0}^{K-1}\langle \nabla J(\boldsymbol{V}_t), \nabla J(\boldsymbol{V}_t) + \boldsymbol{D}_{t,k}\rangle_F .$$

For each k, use the identity

$$\langle \boldsymbol{a}, \boldsymbol{b}\rangle_F = \frac{1}{2}\left(\|\boldsymbol{a}\|_F^2 + \|\boldsymbol{b}\|_F^2 - \|\boldsymbol{a}-\boldsymbol{b}\|_F^2\right)$$

with $\boldsymbol{a} = \nabla J(\boldsymbol{V}_t)$ and $\boldsymbol{b} = \nabla J(\boldsymbol{V}_t) + \boldsymbol{D}_{t,k}$. Since $\boldsymbol{a} - \boldsymbol{b} = -\boldsymbol{D}_{t,k}$, we get

$$\langle \nabla J(\boldsymbol{V}_t), \nabla J(\boldsymbol{V}_t) + \boldsymbol{D}_{t,k}\rangle_F = \frac{1}{2}\left(\|\nabla J(\boldsymbol{V}_t)\|_F^2 + \|\nabla J(\boldsymbol{V}_t) + \boldsymbol{D}_{t,k}\|_F^2 - \|\boldsymbol{D}_{t,k}\|_F^2\right).$$

Dropping the nonpositive term $-\frac{\eta_\ell}{2}\sum_{k=0}^{K-1}\|\nabla J(\boldsymbol{V}_t) + \boldsymbol{D}_{t,k}\|_F^2$ yields

$$\langle \nabla J(\boldsymbol{V}_t), \boldsymbol{\Delta}_t\rangle_F \le -\frac{\eta_\ell K}{2}\|\nabla J(\boldsymbol{V}_t)\|_F^2 + \frac{\eta_\ell}{2}\sum_{k=0}^{K-1}\|\boldsymbol{D}_{t,k}\|_F^2.$$

Substituting this bound into the descent lemma gives the claim. □

A.2 HETEROGENEITY AND LOCAL DRIFT

A.2.1 PROOF OF LEMMA 3.5

Proof. Using the gradient formula,

$$\nabla J_i(\boldsymbol{V}) - \nabla J(\boldsymbol{V}) = 2\,\mathrm{Sym}\left(\boldsymbol{E}_1^\top[\nabla F_i(g(\boldsymbol{V})) - \nabla F(g(\boldsymbol{V}))]\boldsymbol{E}_2^\top\right)\boldsymbol{V}.$$

Since the extractor and symmetrization operators are nonexpansive in the Frobenius norm,

$$\|\nabla J_i(\boldsymbol{V}) - \nabla J(\boldsymbol{V})\|_F \le 2\|\nabla F_i(g(\boldsymbol{V})) - \nabla F(g(\boldsymbol{V}))\|_F\|\boldsymbol{V}\|_F.$$

Averaging over clients and applying the $\boldsymbol{W}$-space heterogeneity assumption proves the result. □

Lemma A.2 (LoRA gradient difference)**.** *For any client i and any $\boldsymbol{U}, \boldsymbol{V}$,*

$$\begin{aligned}\|\nabla J_i(\boldsymbol{U}) - \nabla J_i(\boldsymbol{V})\|_F \le{}& 2L\|\boldsymbol{U}-\boldsymbol{V}\|_F(\|\boldsymbol{U}\|_F + \|\boldsymbol{V}\|_F)\|\boldsymbol{U}\|_F \\ &+ 2\|\nabla F_i(g(\boldsymbol{V}))\|_F\|\boldsymbol{U}-\boldsymbol{V}\|_F. \end{aligned} \tag{16}$$

Consequently, under the local stability condition $\|\boldsymbol{V}_{t,k}^i\|_F \le 2\|\boldsymbol{V}_t\|_F$,

$$\|\nabla J_i(\boldsymbol{V}_{t,k}^i) - \nabla J_i(\boldsymbol{V}_t)\|_F^2 \le \Lambda_t\|\boldsymbol{V}_{t,k}^i - \boldsymbol{V}_t\|_F^2.$$

Proof. Let

$$\boldsymbol{H}_i(\boldsymbol{V}) := \mathrm{Sym}\left(\boldsymbol{E}_1^\top\nabla F_i(g(\boldsymbol{V}))\boldsymbol{E}_2^\top\right).$$

Then $\nabla J_i(\boldsymbol{V}) = 2\boldsymbol{H}_i(\boldsymbol{V})\boldsymbol{V}$, and hence

$$\nabla J_i(\boldsymbol{U}) - \nabla J_i(\boldsymbol{V}) = 2[\boldsymbol{H}_i(\boldsymbol{U}) - \boldsymbol{H}_i(\boldsymbol{V})]\boldsymbol{U} + 2\boldsymbol{H}_i(\boldsymbol{V})(\boldsymbol{U}-\boldsymbol{V}).$$

The second term is bounded by

$$2\|\boldsymbol{H}_i(\boldsymbol{V})\|_F\|\boldsymbol{U}-\boldsymbol{V}\|_F \le 2\|\nabla F_i(g(\boldsymbol{V}))\|_F\|\boldsymbol{U}-\boldsymbol{V}\|_F.$$

For the first term, L-smoothness gives

$$\|\boldsymbol{H}_i(\boldsymbol{U}) - \boldsymbol{H}_i(\boldsymbol{V})\|_F \le L\|g(\boldsymbol{U}) - g(\boldsymbol{V})\|_F.$$

Moreover,

$$\|g(\boldsymbol{U}) - g(\boldsymbol{V})\|_F \le \|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{V}\boldsymbol{V}^\top\|_F \le \|\boldsymbol{U}-\boldsymbol{V}\|_F(\|\boldsymbol{U}\|_F + \|\boldsymbol{V}\|_F).$$

Combining these inequalities proves the first claim.

If $\boldsymbol{U} = \boldsymbol{V}_{t,k}^i$, $\boldsymbol{V} = \boldsymbol{V}_t$, and $\|\boldsymbol{V}_{t,k}^i\|_F \le 2\|\boldsymbol{V}_t\|_F$, then

$$\|\nabla J_i(\boldsymbol{V}_{t,k}^i) - \nabla J_i(\boldsymbol{V}_t)\|_F \le \left(12L\|\boldsymbol{V}_t\|_F^2 + 2\|\nabla F_i(g(\boldsymbol{V}_t))\|_F\right)\|\boldsymbol{V}_{t,k}^i - \boldsymbol{V}_t\|_F.$$

Squaring and using $(a+b)^2 \le 2a^2+2b^2$ yields the stated bound with $\Lambda_t = 288L^2\|\boldsymbol{V}_t\|_F^4 + 8G_t^2$. □

Lemma A.3 (Local stability)**.** *Fix a communication round t and suppose $\|\boldsymbol{V}_t\|_F > 0$. If*

$$4\eta_\ell K G_t \exp(12\eta_\ell K L\|\boldsymbol{V}_t\|_F^2) \le 1,$$

then for every client $i \in [N]$ and every local step $k = 0, \dots, K$,

$$\|\boldsymbol{V}_{t,k}^i - \boldsymbol{V}_t\|_F \le \|\boldsymbol{V}_t\|_F, \qquad \|\boldsymbol{V}_{t,k}^i\|_F \le 2\|\boldsymbol{V}_t\|_F.$$

Proof. Fix a client i and write $d_k = \|\boldsymbol{V}_{t,k}^i - \boldsymbol{V}_t\|_F$. We prove the claim by controlling d_k as long as $d_k \le \|\boldsymbol{V}_t\|_F$. In this regime,

$$\|\boldsymbol{V}_{t,k}^i\|_F \le \|\boldsymbol{V}_t\|_F + \|\boldsymbol{V}_{t,k}^i - \boldsymbol{V}_t\|_F \le 2\|\boldsymbol{V}_t\|_F.$$

Using the LoRA gradient formula and L-smoothness,

$$\begin{aligned}\|\nabla J_i(\boldsymbol{V}_{t,k}^i)\|_F &\le 2\|\nabla F_i(g(\boldsymbol{V}_{t,k}^i))\|_F\|\boldsymbol{V}_{t,k}^i\|_F \\ &\le 2\left(\|\nabla F_i(g(\boldsymbol{V}_t))\|_F + L\|g(\boldsymbol{V}_{t,k}^i) - g(\boldsymbol{V}_t)\|_F\right)2\|\boldsymbol{V}_t\|_F. \end{aligned} \tag{17}$$

Since

$$\|g(\boldsymbol{V}_{t,k}^i) - g(\boldsymbol{V}_t)\|_F \le d_k(\|\boldsymbol{V}_{t,k}^i\|_F + \|\boldsymbol{V}_t\|_F) \le 3\|\boldsymbol{V}_t\|_F d_k,$$

we have

$$\|\nabla J_i(\boldsymbol{V}_{t,k}^i)\|_F \le 4\|\boldsymbol{V}_t\|_F G_t + 12L\|\boldsymbol{V}_t\|_F^2 d_k.$$

Therefore,

$$d_{k+1} \le (1 + 12\eta_\ell L\|\boldsymbol{V}_t\|_F^2)d_k + 4\eta_\ell\|\boldsymbol{V}_t\|_F G_t.$$

Since $d_0 = 0$, recursion gives

$$d_k \le 4\eta_\ell\|\boldsymbol{V}_t\|_F G_t \sum_{s=0}^{k-1}(1 + 12\eta_\ell L\|\boldsymbol{V}_t\|_F^2)^s \le 4\eta_\ell k\|\boldsymbol{V}_t\|_F G_t \exp(12\eta_\ell kL\|\boldsymbol{V}_t\|_F^2).$$

For $k \le K$, the stepsize condition implies $d_k \le \|\boldsymbol{V}_t\|_F$, which closes the induction. The bound $\|\boldsymbol{V}_{t,k}^i\|_F \le 2\|\boldsymbol{V}_t\|_F$ follows by the triangle inequality. □

Remark A.4 (Simpler sufficient stability condition). The condition in Lemma A.3 is implied by

$$12\eta_\ell KL\|\boldsymbol{V}_t\|_F^2 \le 1, \qquad 4e\eta_\ell KG_t \le 1.$$

If $\|\boldsymbol{V}_t\|_F = 0$, then $\boldsymbol{V}_t = \boldsymbol{0}$ and $\nabla J_i(\boldsymbol{0}) = \boldsymbol{0}$ for all i, so all local iterates remain at zero.

A.2.2 Proof of Lemma 3.6

Proof. For this proof, write the mean-squared local drift as

$$A_{t,k} := \frac{1}{N}\sum_{i=1}^N \|\boldsymbol{V}_{t,k}^i - \boldsymbol{V}_t\|_F^2.$$

For $K = 1$, $A_{t,0} = 0$ and the variance identity gives

$$A_{t,1} = \eta_\ell^2 \frac{1}{N}\sum_{i=1}^N \|\nabla J_i(\boldsymbol{V}_t)\|_F^2 \le \eta_\ell^2 \left(\|\nabla J(\boldsymbol{V}_t)\|_F^2 + 4\|\boldsymbol{V}_t\|_F^2\sigma_g^2\right),$$

which implies the stated bound. We now consider $K \ge 2$.

For each client,

$$\begin{aligned}\boldsymbol{V}_{t,k}^i - \boldsymbol{V}_t = \boldsymbol{V}_{t,k-1}^i - \boldsymbol{V}_t - \eta_\ell\left(\nabla J_i(\boldsymbol{V}_{t,k-1}^i) - \nabla J_i(\boldsymbol{V}_t)\right)\\ - \eta_\ell\left(\nabla J_i(\boldsymbol{V}_t) - \nabla J(\boldsymbol{V}_t)\right) - \eta_\ell\nabla J(\boldsymbol{V}_t).\end{aligned} \tag{18}$$

Using

$$\|\boldsymbol{a} + \boldsymbol{b}\|_F^2 \le \left(1 + \frac{1}{2K-1}\right)\|\boldsymbol{a}\|_F^2 + 2K\|\boldsymbol{b}\|_F^2$$

and $\|\boldsymbol{x} + \boldsymbol{y} + \boldsymbol{z}\|_F^2 \le 3(\|\boldsymbol{x}\|_F^2 + \|\boldsymbol{y}\|_F^2 + \|\boldsymbol{z}\|_F^2)$ gives

$$\begin{aligned}A_{t,k} \le \left(1 + \frac{1}{2K-1}\right)A_{t,k-1}\\ + 6K\eta_\ell^2\frac{1}{N}\sum_{i=1}^N\|\nabla J_i(\boldsymbol{V}_{t,k-1}^i) - \nabla J_i(\boldsymbol{V}_t)\|_F^2\\ + 6K\eta_\ell^2\frac{1}{N}\sum_{i=1}^N\|\nabla J_i(\boldsymbol{V}_t) - \nabla J(\boldsymbol{V}_t)\|_F^2 + 6K\eta_\ell^2\|\nabla J(\boldsymbol{V}_t)\|_F^2.\end{aligned} \tag{19}$$

By Lemma A.2,

$$\frac{1}{N}\sum_{i=1}^N\|\nabla J_i(\boldsymbol{V}_{t,k-1}^i) - \nabla J_i(\boldsymbol{V}_t)\|_F^2 \le \Lambda_t A_{t,k-1}.$$

By Lemma 3.5,

$$\frac{1}{N}\sum_{i=1}^N\|\nabla J_i(\boldsymbol{V}_t) - \nabla J(\boldsymbol{V}_t)\|_F^2 \le 4\|\boldsymbol{V}_t\|_F^2\sigma_g^2.$$

Thus

$$A_{t,k} \leq \left(1 + \frac{1}{2K-1} + 6K\eta_\ell^2\Lambda_t\right) A_{t,k-1} + 6K\eta_\ell^2 \left(\|\nabla J(\boldsymbol{V}_t)\|_F^2 + 4\|\boldsymbol{V}_t\|_F^2\sigma_g^2\right).$$

The stepsize condition implies

$$1 + \frac{1}{2K-1} + 6K\eta_\ell^2\Lambda_t \leq 1 + \frac{1}{K-1},$$

for $K \geq 2$. Since $A_{t,0} = 0$, unrolling the recursion yields

$$A_{t,k} \leq 6K\eta_\ell^2 \left(\|\nabla J(\boldsymbol{V}_t)\|_F^2 + 4\|\boldsymbol{V}_t\|_F^2\sigma_g^2\right) \sum_{p=0}^{k-1} \left(1 + \frac{1}{K-1}\right)^p.$$

For $k \leq K$,

$$\sum_{p=0}^{k-1} \left(1 + \frac{1}{K-1}\right)^p \leq 5K,$$

which proves the result. □

Why the max-client gradient is retained. We keep the max-client quantity G_t in the main proof. Replacing it by an average quantity would create weighted terms of the form

$$\frac{1}{N}\sum_{i=1}^{N} \|\nabla F_i(g(\boldsymbol{V}_t))\|_F^2 \|\boldsymbol{V}_{t,k}^i - \boldsymbol{V}_t\|_F^2,$$

which cannot be bounded by a product of two averages without additional assumptions.

A.3 Global update and squared drift penalty bounds

In the proofs below, $A_{t,k} = N^{-1}\sum_i \|\boldsymbol{V}_{t,k}^i - \boldsymbol{V}_t\|_F^2$ is the same mean-squared drift used above.

Lemma A.5 (Global update norm)**.** *Suppose local stability holds and*

$$6K\eta_\ell^2\Lambda_t \leq \frac{1}{2K-1}.$$

Then

$$\|\boldsymbol{\Delta}_t\|_F \leq \eta_\ell K \left[\|\nabla J(\boldsymbol{V}_t)\|_F + \sqrt{30}\, K\eta_\ell\sqrt{\Lambda_t}\sqrt{\|\nabla J(\boldsymbol{V}_t)\|_F^2 + 4\|\boldsymbol{V}_t\|_F^2\sigma_g^2}\right].$$

Proof. Using the aggregation rule and the local updates,

$$\boldsymbol{\Delta}_t = -\eta_\ell \sum_{k=0}^{K-1} \left(\nabla J(\boldsymbol{V}_t) + \boldsymbol{D}_{t,k}\right).$$

Thus

$$\|\boldsymbol{\Delta}_t\|_F \leq \eta_\ell K\|\nabla J(\boldsymbol{V}_t)\|_F + \eta_\ell \sum_{k=0}^{K-1} \|\boldsymbol{D}_{t,k}\|_F.$$

By Jensen's inequality and Lemma A.2,

$$\|\boldsymbol{D}_{t,k}\|_F^2 \leq \Lambda_t A_{t,k}.$$

Applying Lemma 3.6 gives the stated bound. □

Corollary A.6 (Simplified global update bounds)**.** *Under the conditions of Lemma A.5, if*

$$30\eta_\ell^2 K^2\Lambda_t \leq 1,$$

then

$$\|\boldsymbol{\Delta}_t\|_F^2 \leq 4\eta_\ell^2 K^2 \left(\|\nabla J(\boldsymbol{V}_t)\|_F^2 + 4\|\boldsymbol{V}_t\|_F^2\sigma_g^2\right),$$

$$\|\boldsymbol{\Delta}_t\|_F^3 \leq 8\eta_\ell^3 K^3 \left(\|\nabla J(\boldsymbol{V}_t)\|_F^2 + 4\|\boldsymbol{V}_t\|_F^2\sigma_g^2\right)^{3/2},$$

and

$$\|\boldsymbol{\Delta}_t\|_F^4 \leq 16\eta_\ell^4 K^4 \left(\|\nabla J(\boldsymbol{V}_t)\|_F^2 + 4\|\boldsymbol{V}_t\|_F^2\sigma_g^2\right)^2.$$

Proof. Lemma A.5 and $30\eta_\ell^2 K^2 \Lambda_t \le 1$ imply

$$\|\boldsymbol{\Delta}_t\|_F \le 2\eta_\ell K \sqrt{\|\nabla J(\boldsymbol{V}_t)\|_F^2 + 4\|\boldsymbol{V}_t\|_F^2 \sigma_g^2}.$$

The three displayed bounds follow by raising this inequality to the second, third, and fourth powers. □

Lemma A.7 (Bound on the squared drift penalty)**.** *Suppose the conditions of Lemma 3.6 hold. Then*

$$\frac{\eta_\ell}{2} \sum_{k=0}^{K-1} \|\boldsymbol{D}_{t,k}\|_F^2 \le 15K^3\eta_\ell^3\Lambda_t \left(\|\nabla J(\boldsymbol{V}_t)\|_F^2 + 4\|\boldsymbol{V}_t\|_F^2\sigma_g^2\right).$$

Proof. As in Lemma A.5,

$$\|\boldsymbol{D}_{t,k}\|_F^2 \le \Lambda_t A_{t,k}.$$

Using Lemma 3.6,

$$\|\boldsymbol{D}_{t,k}\|_F^2 \le 30K^2\eta_\ell^2\Lambda_t \left(\|\nabla J(\boldsymbol{V}_t)\|_F^2 + 4\|\boldsymbol{V}_t\|_F^2\sigma_g^2\right).$$

This bound is uniform over $k = 0, \dots, K-1$. Summing over k and multiplying by $\eta_\ell/2$ gives the result. □

A.4 One-round descent and the logarithmic rate

A.4.1 Proof of Lemma 3.7

Proof. Suppose for the moment that $h_t^2\Lambda_t \le 1/60$. Then

$$4h_t G_t \exp(12h_t L\|\boldsymbol{V}_t\|_F^2) \le \frac{1}{\sqrt{30}} \exp\left(\frac{1}{\sqrt{120}}\right) < 1,$$

and

$$6K\eta_{\ell,t}^2\Lambda_t = \frac{6h_t^2\Lambda_t}{K} \le \frac{1}{10K} \le \frac{1}{2K-1}.$$

Hence Lemmas A.3 and 3.6 apply, and factor averaging gives $\|\boldsymbol{\Delta}_t\|_F \le \|\boldsymbol{V}_t\|_F$. If $\boldsymbol{V}_t = \boldsymbol{0}$, all local iterates remain at zero and the estimates below hold trivially. Applying Lemma 3.4 to the global objective and using $\|\boldsymbol{\Delta}_t\|_F \le \|\boldsymbol{V}_t\|_F$ gives

$$J(\boldsymbol{V}_{t+1}) \le J(\boldsymbol{V}_t) + \langle \nabla J(\boldsymbol{V}_t), \boldsymbol{\Delta}_t\rangle_F + C_t\|\boldsymbol{\Delta}_t\|_F^2,$$

where

$$C_t = \frac{9\sqrt{2}}{4}L\|\boldsymbol{V}_t\|_F^2 + \|\nabla F(g(\boldsymbol{V}_t))\|_F \le \frac{9\sqrt{2}}{4}L\|\boldsymbol{V}_t\|_F^2 + G_t.$$

By weighted Cauchy–Schwarz,

$$h_t C_t \le \sqrt{\left(\frac{(9\sqrt{2}/4)^2}{288} + \frac{1}{8}\right) h_t^2\Lambda_t} \le \sqrt{\frac{41}{15360}} < \frac{1}{4}.$$

Let $\bar{\boldsymbol{D}}_t = K^{-1}\sum_{k=0}^{K-1}\boldsymbol{D}_{t,k}$. The exact update $\boldsymbol{\Delta}_t = -h_t(\nabla J(\boldsymbol{V}_t) + \bar{\boldsymbol{D}}_t)$ implies

$$\langle \nabla J(\boldsymbol{V}_t), \boldsymbol{\Delta}_t\rangle_F = -\frac{h_t}{2}\|\nabla J(\boldsymbol{V}_t)\|_F^2 - \frac{\|\boldsymbol{\Delta}_t\|_F^2}{2h_t} + \frac{h_t}{2}\|\bar{\boldsymbol{D}}_t\|_F^2.$$

Jensen's inequality and Lemma A.7 give

$$\frac{h_t}{2}\|\bar{\boldsymbol{D}}_t\|_F^2 \le \frac{\eta_{\ell,t}}{2}\sum_{k=0}^{K-1}\|\boldsymbol{D}_{t,k}\|_F^2 \le 15h_t^3\Lambda_t\left(\|\nabla J(\boldsymbol{V}_t)\|_F^2 + 4\|\boldsymbol{V}_t\|_F^2\sigma_g^2\right).$$

Since $15h_t^2\Lambda_t \le 1/4$, it follows that

$$J(\boldsymbol{V}_{t+1}) \le J(\boldsymbol{V}_t) - \frac{h_t}{4}\|\nabla J(\boldsymbol{V}_t)\|_F^2 - \frac{\|\boldsymbol{\Delta}_t\|_F^2}{4h_t} + 60h_t^3\Lambda_t\|\boldsymbol{V}_t\|_F^2\sigma_g^2. \tag{20}$$

□

A.4.2 Complete statement and proof of Theorem 3.8

Theorem 3.8 (Complete statement)**.** *Suppose each F_i is L-smooth, $J \geq J^\star$, and the $\boldsymbol{W}$-space heterogeneity bound holds with constant σ_g^2 at all generated server iterates. Fix $K \geq 1$ and $T \geq 1$. Consider full-batch local gradient descent with full participation and factor averaging. Define the initial quantities*

$$R_0 = \|\boldsymbol{V}_0\|_F, \qquad G_0 = \max_{i\in[N]} \|\nabla F_i(g(\boldsymbol{V}_0))\|_F, \qquad E_0 = J(\boldsymbol{V}_0) - J^\star,$$

and the constants

$$M_V = 2R_0^2 + 8(E_0 + 1),$$
$$M_G = G_0 + L(M_V + R_0^2), \qquad M_\Lambda = 288L^2M_V^2 + 8M_G^2.$$

If $M_\Lambda > 0$, choose

$$0 < \gamma_0 \leq \min\left\{1, \frac{1}{\sqrt{60M_\Lambda}}, \frac{1}{(1020M_\Lambda M_V)^{1/3}}\right\}, \qquad \gamma = \frac{\gamma_0}{1+\sigma_g},$$

and set

$$\eta_{\ell,t}K = \frac{\gamma}{t+1}, \qquad t = 0, \ldots, T-1.$$

Then

$$\min_{0\leq t<T} \|\nabla J(\boldsymbol{V}_t)\|_F^2 \leq \frac{4(E_0+1)}{\gamma \log(T+1)} = \frac{4(E_0+1)(1+\sigma_g)}{\gamma_0 \log(T+1)}.$$

Thus the rate is $O((1+\sigma_g)/\log(T+1))$ when $L, R_0, G_0, E_0, \gamma_0$ are held fixed. No uniform trajectory bound is assumed. If $M_\Lambda = 0$, then $L = G_0 = 0$, all local gradients vanish, and the result is immediate for any positive harmonic prefactor.

Proof. Write $h_t = \eta_{\ell,t}K$ and $\mathcal{H}_t = \sum_{j=1}^t j^{-1}$, with $\mathcal{H}_0 = 0$. We first retain a negative update-norm term in the descent estimate, then control the trajectory and the accumulated residual by induction.

A refined one-round estimate. See the proof for Lemma 3.7.

Joint control of the trajectory and residual. Define

$$r_t = 60h_t^3\Lambda_t\|\boldsymbol{V}_t\|_F^2\sigma_g^2, \qquad \mathcal{R}_t = \sum_{s=0}^{t-1} r_s.$$

We prove simultaneously that $\mathcal{R}_t \leq 1$ for all t and that (20) is valid at every round. Suppose it is valid for rounds $0, \ldots, t-1$ and $\mathcal{R}_s \leq 1$ for $0 \leq s \leq t$. The base case $t = 0$ has empty sums. Summing the preceding descent inequalities yields

$$J(\boldsymbol{V}_t) + \frac{1}{4}\sum_{s<t} h_s\|\nabla J(\boldsymbol{V}_s)\|_F^2 + \frac{1}{4}\sum_{s<t}\frac{\|\boldsymbol{\Delta}_s\|_F^2}{h_s} \leq J(\boldsymbol{V}_0) + \mathcal{R}_t. \tag{21}$$

Using $J \geq J^\star$ gives

$$\sum_{s<t}\frac{\|\boldsymbol{\Delta}_s\|_F^2}{h_s} \leq 4(E_0 + \mathcal{R}_t) \leq 4(E_0+1).$$

Consequently, weighted Cauchy–Schwarz implies

$$\begin{aligned}\|\boldsymbol{V}_t\|_F^2 &\leq 2R_0^2 + 2\left(\sum_{s<t} h_s\right)\left(\sum_{s<t}\frac{\|\boldsymbol{\Delta}_s\|_F^2}{h_s}\right)\\ &\leq 2R_0^2 + 8\gamma(E_0+1)\mathcal{H}_t \leq M_V(1+\mathcal{H}_t),\end{aligned}$$

where $\gamma \leq 1$ was used. Smoothness then gives

$$\begin{aligned}G_t &\leq G_0 + L\|g(\boldsymbol{V}_t) - g(\boldsymbol{V}_0)\|_F\\ &\leq G_0 + L(\|\boldsymbol{V}_t\|_F^2 + R_0^2) \leq M_G(1+\mathcal{H}_t).\end{aligned}$$

Thus $\Lambda_t \le M_\Lambda(1+\mathcal{H}_t)^2$. Since $\mathcal{H}_t \le t$, the stepsize condition at round t follows:

$$h_t^2\Lambda_t \le \gamma^2 M_\Lambda \frac{(1+\mathcal{H}_t)^2}{(t+1)^2} \le \gamma_0^2 M_\Lambda \le \frac{1}{60}.$$

Therefore (20) also holds at round t. The established bounds for every $s \le t$ imply

$$r_s \le 60\gamma^3\sigma_g^2 M_\Lambda M_V \frac{(1+\mathcal{H}_s)^3}{(s+1)^3}.$$

Moreover,

$$\begin{aligned}\sum_{s=0}^{\infty} \frac{(1+\mathcal{H}_s)^3}{(s+1)^3} &\le \sum_{n=1}^{\infty} \frac{(2+\log n)^3}{n^3} \\ &\le 8 + \int_1^{\infty} \frac{(2+\log x)^3}{x^3}\,dx = \frac{135}{8} < 17,\end{aligned}$$

where the sum–integral comparison uses the decreasing integrand. It follows that

$$\mathcal{R}_{t+1} \le 1020\gamma^3\sigma_g^2 M_\Lambda M_V = 1020\gamma_0^3 M_\Lambda M_V \frac{\sigma_g^2}{(1+\sigma_g)^3} \le 1.$$

This closes the induction. In particular, the trajectory estimates are derived rather than assumed, and all constants are independent of T.

Telescoping and the rate. By (21) at $t = T$ and $J(\boldsymbol{V}_T) \ge J^\star$,

$$\sum_{t=0}^{T-1} h_t \|\nabla J(\boldsymbol{V}_t)\|_F^2 \le 4(E_0+1).$$

Finally,

$$\sum_{t=0}^{T-1} h_t = \gamma \sum_{t=0}^{T-1} \frac{1}{t+1} \ge \gamma \log(T+1).$$

Bounding the minimum squared gradient norm by this weighted average proves the claim. □

Explicit dependence on local steps and heterogeneity. Under $\eta_{\ell,t} = \gamma/[K(t+1)]$, every explicit occurrence of K in the convergence bound is through $\eta_{\ell,t}K$. The $1/K$ normalization therefore cancels this dependence; the theorem does not provide an additional $1/K$ improvement per communication round. The choice $\gamma = \gamma_0/(1+\sigma_g)$ makes the heterogeneity dependence explicit. The hidden constants depend on L, R_0, G_0, E_0 and γ_0. Joint comparisons across K or σ_g require these initial quantities and an applicable heterogeneity bound to remain uniformly controlled.

A.5 Complete statement and proof of Theorem 3.10

Theorem 3.10 (Complete statement)**.** *Suppose each F_i is L-smooth, the $\boldsymbol{W}$-space heterogeneity bound holds with constant σ_g^2, and $J \geq J^\star$. Fix $K \geq 1$ and $T \geq 1$. Consider full-batch local gradient descent with full participation, factor averaging, and a common local stepsize $\eta_\ell > 0$ that is constant across all T rounds. Suppose that along this trajectory, for every $0 \leq t < T$,*

$$\|\boldsymbol{V}_t\|_F^2 \leq B_V, \qquad G_t^2 = \max_{i\in[N]} \|\nabla F_i(g(\boldsymbol{V}_t))\|_F^2 \leq B_G^2.$$

Choose $\gamma > 0$ satisfying

$$12\gamma L B_V \leq 1, \qquad 4e\gamma B_G \leq 1,$$
$$60\gamma^2(288L^2B_V^2 + 8B_G^2) \leq 1, \quad 2\gamma\left(\frac{9\sqrt{2}}{4}LB_V + B_G\right) \leq 1.$$

If $0 < \eta_\ell K \leq \gamma$, then

$$\min_{0\leq t<T} \|\nabla J(\boldsymbol{V}_t)\|_F^2 \leq \frac{4\big(J(\boldsymbol{V}_0) - J^\star\big)}{\eta_\ell K T} + 240(\eta_\ell K)^2(288L^2B_V^2 + 8B_G^2)B_V\sigma_g^2.$$

In particular, choosing $\eta_\ell = \gamma/(KT^{1/3})$ gives

$$\min_{0\leq t<T} \|\nabla J(\boldsymbol{V}_t)\|_F^2 \leq \left[\frac{4\big(J(\boldsymbol{V}_0) - J^\star\big)}{\gamma} + 240\gamma^2(288L^2B_V^2 + 8B_G^2)B_V\sigma_g^2\right] T^{-2/3}.$$

Thus the squared-gradient stationarity rate is $O(T^{-2/3})$ when K, L, B_V, B_G, σ_g, γ, and the initial objective gap are independent of T, with the trajectory bounds holding uniformly for the runs generated by this stepsize choice.

For fixed γ, this is $O((1+\sigma_g^2)T^{-2/3})$ when the remaining quantities are uniformly controlled. Alternatively, choose

$$\eta_\ell = \frac{\gamma}{K(1+\sigma_g^2)^{1/3}T^{1/3}}$$

with the same admissible γ. If the stated trajectory bounds hold for the runs generated by this choice, then

$$\min_{0\leq t<T} \|\nabla J(\boldsymbol{V}_t)\|_F^2 \leq \left[\frac{4\big(J(\boldsymbol{V}_0) - J^\star\big)}{\gamma} + 240\gamma^2(288L^2B_V^2 + 8B_G^2)B_V\right] \cdot \frac{(1+\sigma_g^2)^{1/3}}{T^{2/3}}.$$

The two stepsize choices cancel the explicit K factor through $\eta_\ell K$. For dependence jointly in T, K, σ_g, the bounds B_V, B_G, the constants L, γ, and the initial objective gap must remain uniformly controlled over the corresponding runs.

Proof. The trajectory bounds imply

$$\|\nabla F(g(\boldsymbol{V}_t))\|_F \leq \frac{1}{N}\sum_{i=1}^N \|\nabla F_i(g(\boldsymbol{V}_t))\|_F \leq B_G, \qquad \Lambda_t \leq 288L^2B_V^2 + 8B_G^2.$$

The first two restrictions on γ, together with $\eta_\ell K \leq \gamma$, imply the sufficient stability conditions in Remark A.4. Hence Lemma A.3 gives $\|\boldsymbol{V}_{t,k}^i - \boldsymbol{V}_t\|_F \leq \|\boldsymbol{V}_t\|_F$ for all i and $0 \leq k \leq K$. When $\boldsymbol{V}_t = \boldsymbol{0}$, all client gradients in factor space vanish and the same statement holds directly. Factor averaging therefore yields

$$\|\boldsymbol{\Delta}_t\|_F = \left\|\frac{1}{N}\sum_{i=1}^N(\boldsymbol{V}_{t,K}^i - \boldsymbol{V}_t)\right\|_F \leq \frac{1}{N}\sum_{i=1}^N \|\boldsymbol{V}_{t,K}^i - \boldsymbol{V}_t\|_F \leq \|\boldsymbol{V}_t\|_F.$$

Moreover, the third restriction on γ gives

$$6K\eta_\ell^2\Lambda_t \le \frac{6\gamma^2}{K}(288L^2B_V^2 + 8B_G^2) \le \frac{1}{10K} \le \frac{1}{2K-1}.$$

Thus Lemmas 3.6 and A.7 apply.

Using $\|\boldsymbol{\Delta}_t\|_F \le \|\boldsymbol{V}_t\|_F$, the higher-order terms in Lemma 3.4, applied to the global objective, are bounded as follows:

$$\begin{aligned}&\sqrt{2}L\|\boldsymbol{\Delta}_t\|_F^2\|\boldsymbol{V}_t\|_F^2 + \sqrt{2}L\|\boldsymbol{\Delta}_t\|_F^3\|\boldsymbol{V}_t\|_F + \frac{\sqrt{2}L}{4}\|\boldsymbol{\Delta}_t\|_F^4 + \|\nabla F(g(\boldsymbol{V}_t))\|_F\|\boldsymbol{\Delta}_t\|_F^2 \\ &\quad\le \left(\frac{9\sqrt{2}}{4}LB_V + B_G\right)\|\boldsymbol{\Delta}_t\|_F^2.\end{aligned}$$

The update identity and Cauchy–Schwarz give

$$\boldsymbol{\Delta}_t = -\eta_\ell \sum_{k=0}^{K-1}\big(\nabla J(\boldsymbol{V}_t) + \boldsymbol{D}_{t,k}\big), \qquad \|\boldsymbol{\Delta}_t\|_F^2 \le \eta_\ell^2 K \sum_{k=0}^{K-1}\|\nabla J(\boldsymbol{V}_t) + \boldsymbol{D}_{t,k}\|_F^2.$$

Retaining the full inner-product identity from the proof of Lemma A.1, we have

$$\begin{aligned}\langle \nabla J(\boldsymbol{V}_t), \boldsymbol{\Delta}_t\rangle_F = &-\frac{\eta_\ell K}{2}\|\nabla J(\boldsymbol{V}_t)\|_F^2 - \frac{\eta_\ell}{2}\sum_{k=0}^{K-1}\|\nabla J(\boldsymbol{V}_t) + \boldsymbol{D}_{t,k}\|_F^2 \\ &+ \frac{\eta_\ell}{2}\sum_{k=0}^{K-1}\|\boldsymbol{D}_{t,k}\|_F^2.\end{aligned}$$

The fourth restriction on γ ensures that

$$\eta_\ell^2 K\left(\frac{9\sqrt{2}}{4}LB_V + B_G\right) \le \frac{\eta_\ell}{2}.$$

Consequently, the negative sum in the inner-product identity absorbs all the higher-order terms, leaving

$$J(\boldsymbol{V}_{t+1}) \le J(\boldsymbol{V}_t) - \frac{\eta_\ell K}{2}\|\nabla J(\boldsymbol{V}_t)\|_F^2 + \frac{\eta_\ell}{2}\sum_{k=0}^{K-1}\|\boldsymbol{D}_{t,k}\|_F^2.$$

By Lemma A.7,

$$\begin{aligned}J(\boldsymbol{V}_{t+1}) \le J(\boldsymbol{V}_t) &- \frac{\eta_\ell K}{2}\|\nabla J(\boldsymbol{V}_t)\|_F^2 \\ &+ 15(\eta_\ell K)^3\Lambda_t\big(\|\nabla J(\boldsymbol{V}_t)\|_F^2 + 4\|\boldsymbol{V}_t\|_F^2\sigma_g^2\big).\end{aligned}$$

Since $15(\eta_\ell K)^2\Lambda_t \le 1/4$, we obtain

$$\begin{aligned}J(\boldsymbol{V}_{t+1}) \le J(\boldsymbol{V}_t) &- \frac{\eta_\ell K}{4}\|\nabla J(\boldsymbol{V}_t)\|_F^2 \\ &+ 60(\eta_\ell K)^3(288L^2B_V^2 + 8B_G^2)B_V\sigma_g^2.\end{aligned}$$

Summing over $t = 0, \dots, T-1$ and using $J(\boldsymbol{V}_T) \ge J^\star$ gives

$$\begin{aligned}\frac{\eta_\ell K}{4}\sum_{t=0}^{T-1}\|\nabla J(\boldsymbol{V}_t)\|_F^2 \le J(\boldsymbol{V}_0) - J^\star \\ + 60T(\eta_\ell K)^3(288L^2B_V^2 + 8B_G^2)B_V\sigma_g^2.\end{aligned}$$

Dividing by $\eta_\ell KT/4$ and bounding the minimum by the average proves the constant-stepsize bound. Finally, $\eta_\ell K = \gamma T^{-1/3}$ makes both terms proportional to $T^{-2/3}$, yielding the claimed rate. □

Remark A.8 (Stepsize conditions for Theorem 3.10). Under the trajectory bounds in Theorem 3.10, the four restrictions on $\gamma > 0$ are equivalent to the single condition

$$60\gamma^2(288L^2B_V^2 + 8B_G^2) \le 1.$$

Indeed, this condition implies

$$12\gamma LB_V \le \frac{1}{\sqrt{120}} < 1, \qquad 4e\gamma B_G \le \frac{e}{\sqrt{30}} < 1,$$

and weighted Cauchy–Schwarz gives

$$\begin{aligned} 2\gamma\left(\frac{9\sqrt{2}}{4}LB_V + B_G\right) &\le \gamma\sqrt{\frac{41}{64}(288L^2B_V^2 + 8B_G^2)} \\ &\le \sqrt{\frac{41}{3840}} < 1. \end{aligned}$$

Thus the third restriction stated above already implies the other three, without further reducing the admissible stepsize range. For $288L^2B_V^2 + 8B_G^2 > 0$, the equivalent explicit upper bound is

$$0 < \gamma \le \frac{1}{\sqrt{60(288L^2B_V^2 + 8B_G^2)}}.$$

Any fixed $0 < \eta_\ell K \le \gamma$ gives the constant-stepsize bound stated above. Taking γ at this upper bound yields the feasible horizon-dependent choice

$$\eta_\ell = \frac{1}{KT^{1/3}\sqrt{60(288L^2B_V^2 + 8B_G^2)}},$$

which remains constant across rounds within each run. Substitution into the explicit bound gives

$$\begin{aligned} \min_{0\le t<T} \|\nabla J(\boldsymbol{V}_t)\|_F^2 \le \Big[&4\sqrt{60(288L^2B_V^2 + 8B_G^2)}\big(J(\boldsymbol{V}_0) - J^\star\big) \\ &+ 4B_V\sigma_g^2\Big]T^{-2/3}. \end{aligned}$$

This choice need not minimize the convergence upper bound. If $288L^2B_V^2 + 8B_G^2 = 0$, the four inequalities impose no restriction on γ; the reciprocal formulas above apply only when their denominator is positive.

The stepsize depends on the factor-norm bound B_V and the maximum client gradient bound B_G in $\boldsymbol{W}$-space. No additional ratio involving the current global factor-space gradient is required: the condition $15(\eta_\ell K)^2\Lambda_t \le 1/4$ absorbs the gradient-dependent drift term for every gradient norm, including zero. The proof retains the cubic heterogeneity residual and does not guarantee objective decrease at every round. The trajectory bounds remain assumptions on the iterates generated by the chosen stepsize; for the stated $O(T^{-2/3})$ rate, their constants and the initial objective gap must remain uniform in T.

Explicit dependence on local steps and heterogeneity. For a general admissible constant stepsize, the finite-round bound contains K through $1/(\eta_\ell KT)$ and $(\eta_\ell K)^2\sigma_g^2$. Holding the trajectory bounds fixed, increasing K at a fixed local stepsize decreases the first term and increases the second, subject to the stepsize restrictions. Under $\eta_\ell = \gamma/(KT^{1/3})$, the effective stepsize is $\eta_\ell K = \gamma T^{-1/3}$, so K cancels from the displayed bound. This does not establish a K-fold improvement per communication round; different K can still change the trajectory bounds and the admissible γ.

For the alternative stepsize,

$$\eta_\ell K = \frac{\gamma}{(1+\sigma_g^2)^{1/3}T^{1/3}} \le \gamma,$$

so it is also admissible under the same uniform trajectory bounds. Substitution into the general finite-round bound gives

$$\begin{aligned} \min_{0\le t<T} \|\nabla J(\boldsymbol{V}_t)\|_F^2 \le \Bigg[&\frac{4\big(J(\boldsymbol{V}_0) - J^\star\big)}{\gamma}(1+\sigma_g^2)^{1/3} \\ &+ 240\gamma^2(288L^2B_V^2 + 8B_G^2)B_V\frac{\sigma_g^2}{(1+\sigma_g^2)^{2/3}}\Bigg]T^{-2/3}. \end{aligned}$$

Using $\sigma_g^2/(1+\sigma_g^2)^{2/3} \le (1+\sigma_g^2)^{1/3}$ proves the alternative bound in the complete statement. It exposes a cube-root dependence on the heterogeneity variance σ_g^2 after adjusting the stepsize; the original choice with fixed γ has a linear dependence on that variance. Neither statement removes possible implicit dependence through the trajectory bounds. Joint comparisons require L, B_V, B_G, γ and the initial objective gap to remain uniformly controlled. The alternative choice remains constant across rounds within each run and preserves the $T^{-2/3}$ exponent.

B Additional Experiments

B.1 CIFAR-10 experimental details

Data and preprocessing. Both CIFAR-10 experiments use a stratified split of 45,000 training and 5,000 validation examples. The official 10,000-example test set is reserved for final evaluation. Images are resized to 224×224 and normalized using ImageNet statistics, without data augmentation.

Client partitions. We distribute the training examples among 100 clients. The IID partition assigns 450 randomly shuffled examples to each client. Non-IID partitions use class-specific proportions from a symmetric Dirichlet distribution with $\alpha_{\mathrm{Dir}} \in \{0.1, 0.5\}$, followed by multinomial allocation. Client dataset sizes may differ, and the partition is resampled if any client receives no examples.

LoRA parameterization. Both models use rank $r = 8$ and scaling parameter $\alpha_{\mathrm{LoRA}} = 8$, with effective weight

$$\boldsymbol{W} = \boldsymbol{W}_0 + \frac{\alpha_{\mathrm{LoRA}}}{r}\boldsymbol{B}\boldsymbol{A}. \tag{22}$$

The base weight $\boldsymbol{W}_0$ remains frozen; its value and the adapted modules differ between the two experiments as described below.

Multinomial logistic regression. An ImageNet-1K-pretrained ResNet-18 serves as a frozen feature extractor. We use its ℓ_2-normalized, 512-dimensional penultimate-layer representations as inputs to a multinomial logistic-regression classifier. The classifier uses $\boldsymbol{W}_0 = \boldsymbol{0}$. We initialize $\boldsymbol{A}$ using Kaiming initialization and $\boldsymbol{B}$ to zero, and jointly optimize both factors and the classifier bias. LoRA dropout is disabled.

Convolutional ResNet-18 adaptation. We apply LoRA to all 20 convolutional layers of an ImageNet-pretrained ResNet-18, including the input stem and projection shortcuts. Each convolutional weight is flattened into a matrix before applying the parameterization above. Here, $\boldsymbol{W}_0$ contains the pretrained convolutional weights. The LoRA factors and classification head are trained, while the base weights and BatchNorm layers remain frozen.

Local training and aggregation. Both experiments use cross-entropy loss and minibatch SGD. Each selected client starts from the current global model and performs $K = 10$ local steps with batch size 32. Unless otherwise specified, momentum, weight decay, and gradient clipping are disabled. The server applies sample-weighted FedAvg with server learning rate 1, averaging the two LoRA factors separately along with the other trainable parameters.

Participation and evaluation. Under full participation, all 100 clients participate in each of 200 communication rounds. Under partial participation, 10 clients are sampled uniformly without replacement in each of 500 rounds. Training and validation loss and accuracy are evaluated after every round. Test performance is evaluated only after the final round.

Learning-rate and initialization sensitivity. For logistic regression, we sweep the local learning rate over $\{0.03, 0.1, 0.3, 1, 3\}$. To study scalar scale sensitivity, we rescale the initial factors as

$$\boldsymbol{A} \leftarrow c\boldsymbol{A}, \qquad \boldsymbol{B} \leftarrow \boldsymbol{B}/c, \qquad c \in \{0.5, 1, 2, 4\}. \tag{23}$$

This transformation preserves the initial matrix product $\boldsymbol{B}\boldsymbol{A}$.

B.2 Federated commonsense reasoning with Qwen2.5 details

Tasks and data. We construct a five-task benchmark from ARC-Easy, ARC-Challenge, BoolQ, HellaSwag, and PIQA. From each task, we sample 1,000 examples from the official training split and 200 from the validation split. This yields 5,000 training examples and 1,000 held-out evaluation examples. All tasks use a unified multiple-choice format in which the target is the letter of the correct answer.

Client partitions. We simulate 10 clients, each with exactly 500 training examples. Client task proportions are sampled using a symmetric Dirichlet distribution with $\alpha_{\mathrm{Dir}} = 0.1$. The allocation is constrained to preserve both the 500-example client totals and the 1,000-example global total for each task. Data partitioning and client sampling use random seed 42.

Model and adapters. We use Qwen2.5-1.5B-Instruct as the frozen backbone. LoRA adapters are inserted into the query and value projections (`q_proj` and `v_proj`) of every Transformer block. We set $r = 16$, $\alpha_{\mathrm{LoRA}} = 16$, and LoRA dropout to zero. Both factors are optimized while all backbone parameters remain frozen. The resulting model has approximately 2.18 million trainable parameters out of 1.546 billion total parameters, or 0.141%.

Local optimization. Each selected client initializes its adapter from the current global adapter and performs 20 AdamW steps. The per-device batch size is 1, with four-step gradient accumulation, giving effective batch size 4 and 80 processed training examples per client participation. The learning rate is fixed at 2×10^{-4}. AdamW uses $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, and zero weight decay. Gradients are clipped to norm 1.0. Local optimizer states are reinitialized for each client participation.

Participation and aggregation. Training lasts 50 communication rounds. In each round, two of the 10 clients are sampled uniformly without replacement. The server directly and uniformly averages their $\boldsymbol{A}$ and $\boldsymbol{B}$ factors separately, with server learning rate 1. Since all clients have the same dataset size, uniform and sample-size-weighted aggregation are equivalent.

Training objective and numerical settings. Training uses bfloat16, maximum sequence length 512, and gradient checkpointing. Prompt tokens are masked from the objective. The causal language-modeling loss is computed only over the answer letter and the end-of-sequence token.

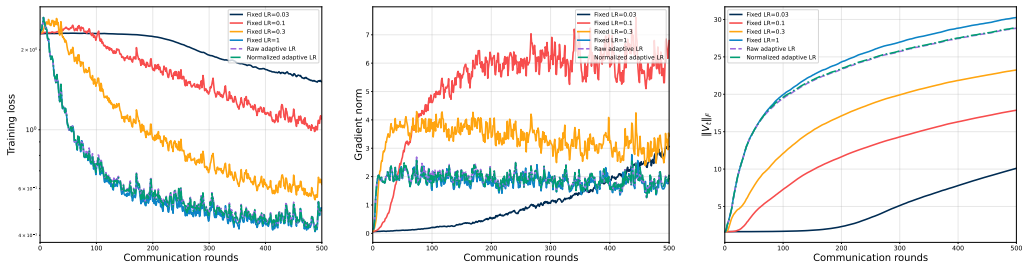
Evaluation. We evaluate the global model every 10 rounds using the 200 held-out examples from each task. For each question, we compute the conditional log-likelihood of every candidate answer letter and select the highest-scoring candidate. We report per-task accuracy, macro accuracy with equal weight for each task, and the minimum accuracy across the five tasks (worst-task accuracy). Under the same protocol, the unadapted backbone obtains 73.8% macro accuracy and 60.5% worst-task accuracy.

Optimization diagnostics. In addition to accuracy and local training loss, we record squared global LoRA factor norms, client factor-update norms, and the norm of the induced $\boldsymbol{BA}$ update. We measure heterogeneity separately for the $\boldsymbol{A}$- and $\boldsymbol{B}$-factor gradients every five rounds using all 10 clients.

B.3 Additional Results

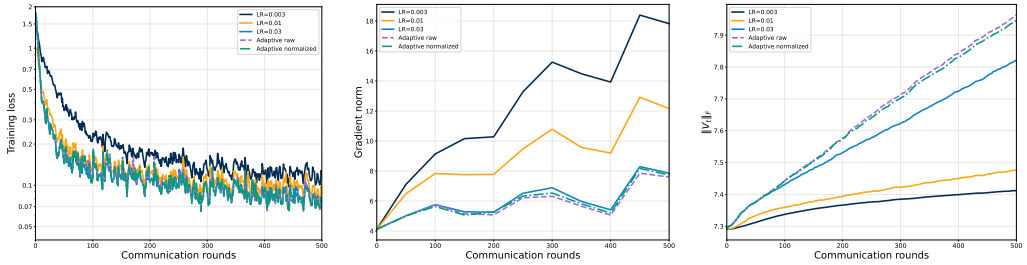
Logistic regression under partial participation (Figure 5) The fixed $\eta_\ell = 1$ baseline and both adaptive schedules reduced training loss substantially faster than the smaller fixed rates, consistent with the learning-rate sensitivity observed under full participation. These low-loss runs also exhibited larger factor norms, reinforcing the need to interpret norm diagnostics alongside training performance.

ResNet-18 under partial participation (Figure 6). Both adaptive schedules achieved training losses close to the fixed $\eta_\ell = 0.03$ baseline, whereas smaller fixed rates reduced loss more slowly. Factor norms increased throughout training, consistent with the main-text observation that smaller factor norms alone did not identify the lowest-loss runs.



(a) Training loss (b) Gradient norm (c) $\boldsymbol{V}$ norm

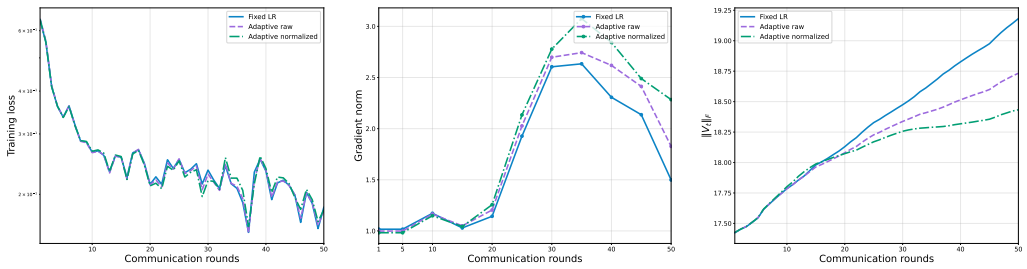
Figure 5: Partial participation results of logistic regression experiment.



(a) Training loss (b) Gradient norm (c) $\boldsymbol{V}$ norm

Figure 6: Partial participation results of fine-tuning ResNet-18 experiment.

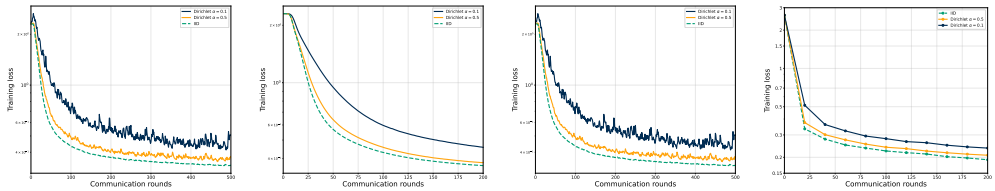
LLM fine-tuning under partial participation (Figure 7). The three schedules produced similar training-loss trajectories over 50 rounds. The adaptive schedules showed slower factor-norm growth but larger gradient diagnostics in later rounds than the fixed-rate baseline. These differences did not translate into a clear training-loss advantage, consistent with the setting-dependent benefits discussed in the main text.



(a) Training loss (b) Gradient norm (c) $\boldsymbol{V}$ norm

Figure 7: Partial participation results of Qwen2.5 1.5B Instruct fine-tuning.

Effect of client heterogeneity (Figure 8). Across the displayed comparisons, the more heterogeneous partition ($\alpha_{\mathrm{Dir}} = 0.1$) yielded higher training losses than the $\alpha_{\mathrm{Dir}} = 0.5$ and i.i.d. partitions. These loss differences complement the accuracy comparisons in Table 2, which also show poorer performance on more heterogeneous partitions.



(a) Partial logistic (b) Full logistic (c) Partial ResNet-18 (d) Full ResNet-18

Figure 8: Effect of client heterogeneity on training loss. Logistic regression and ResNet-18 are evaluated under full and partial client participation.