PROLORA: BIDIRECTIONAL COMMUNICATION-EFFICIENT FEDERATED FINE-TUNING WITH LORA-AWARE ERROR FEEDBACK

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated fine-tuning of large language models (LLMs) using low-rank adaptation (LoRA) addresses data scarcity while preserving data privacy. However, the frequent exchange of LoRA updates leads to bidirectional communication bottlenecks. Existing solutions either reduce communication by restricting the trainable LoRA factors, leading to performance degradation, or compress LoRA updates, resulting in high downlink communication overhead or additional server-side compression. To overcome these challenges, we propose ProLoRA by designing novel and efficient random orthogonal projection (ROP) compressor and LoRA-aware error feedback (LA-EF), which enables efficient bidirectional communication while maintaining high model performance. Specifically, ROP enables direct aggregation of compressed LoRA updates through its linearity, while LoRA-aware error feedback exploits shared LoRA geometry to regulate historical compression residuals and improve model recovery. We theoretically prove that ProLoRA achieves the same $\mathcal{O}(1/\sqrt{T})$ convergence order as full-precision FedAvg. We conduct extensive experiments on multiple datasets and LLMs to demonstrate that ProLoRA reduces total communication costs by up to 96.9% without sacrificing model performance when compared with state-of-the-art (SOTA) solutions.

1 INTRODUCTION

Pre-trained on massive high-quality public data, large language models (LLMs) have demonstrated remarkable performance across a wide range of general tasks. To adapt these models to specific downstream tasks, fine-tuning is essential (Wu et al., 2025; Mao et al., 2025). However, centralized fine-tuning on a single client often fails to provide satisfactory model performance due to limited local data. Federated low-rank adaptation (LoRA) (Hu et al., 2021) fine-tuning (Zhang et al., 2024; Ye et al., 2024), which integrates LoRA-based fine-tuning of LLMs with federated learning (FL) (McMahan et al., 2017), has emerged as a highly promising solution to mitigate this issue. Nevertheless, due to the enormous scale of LLMs, frequent uploading and downloading of LoRA updates between clients and the server incur substantial communication overheads. This makes communication latency the primary bottleneck of the training process, significantly reducing training efficiency (Woisetschläger et al., 2024; Li et al., 2025).

To address this issue, some works have made preliminary explorations. For example, FFA-LoRA (Sun et al., 2024), RoLoRA (Chen et al., 2025), and Fed-SB (Singhal et al., 2025) reduce communication costs by restricting the LoRA parameters updated in each round. Although these methods improve communication efficiency, they may also limit the adaptation capacity of LoRA and thus degrade model performance. Another line of work reduces communication costs by compressing LoRA updates. Specifically, FLASC (Kuo et al., 2026) adopts TopK compression and transmits only LoRA updates with large magnitudes. EcoLoRA (Liu et al., 2025) combines round-robin segmented uploading with adaptive TopK and transmits sparse updates from only part of the LoRA modules in each round. FedSRD (Yan et al., 2026) further exploits the two-factor structure of LoRA to estimate the importance of LoRA updates and transmits only high-importance update elements. However, different clients may retain different LoRA update entries, causing the aggregated LoRA update to become much denser. The server must then either transmit the denser aggregated update, resulting in

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.

high downlink communication overhead, or compress it again, introducing additional compression errors.

To address these challenges, we propose ProLoRA by designing new random orthogonal projection (ROP) compressor and LoRA-aware error feedback (LA-EF), enabling efficient bidirectional communication while preserving full LoRA training and model performance. Specifically, ROP reduces memory and computational overhead by reformulating random projection in matrix form and orthogonalizing the projection bases, while exploiting its linearity to directly aggregate and broadcast compressed updates without server-side recompression. Meanwhile, ROP remains a biased compressor and thus introduces information loss, motivating us to employ error feedback (EF) (Karimireddy et al., 2019; Li & Li, 2023) to compensate for historical projection residuals. However, when standard EF is combined with ROP, it uniformly feeds accumulated historical residuals back into subsequent updates, even though the effect of a residual in one LoRA factor depends on the state of the other factor. Such uniform residual feedback may hinder model performance recovery under high compression and data heterogeneity. To address this issue, we design LA-EF, which uses Gram matrices of the current global LoRA factors to regulate historical residual feedback and improve model performance recovery without additional communication overhead. Our main contributions are summarized as follows.

- We design a novel ROP compressor for LoRA, which reduces computation and memory overheads through matrix projection and orthogonal reconstruction, while exploiting linearity to directly aggregate and broadcast compressed updates without server-side recompression.
- We design a novel LA-EF mechanism, which uses the partner LoRA factors to regulate historical residuals and reduce their projection noise, improving model recovery under high compression and data heterogeneity without additional communication.
- We provide a non-convex convergence analysis for ProLoRA that accounts for time-varying LoRA geometry and historical residual filtering, showing that ProLoRA matches the $\mathcal{O}(1/\sqrt{T})$ convergence order of full-precision FedAvg.
- We conduct extensive experiments on multiple datasets and LLMs to demonstrate that ProLoRA significantly outperforms various baselines in terms of communication efficiency, reducing total communication overheads by up to 96.9% while maintaining model performance.

2 Related Work

Federated LoRA fine-tuning has emerged as a prominent paradigm for distributed fine-tuning of LLMs (Yang et al., 2024). FedIT (Zhang et al., 2024) combines LoRA with FedAvg so that only LoRA modules need to be uploaded and downloaded. However, frequent exchange of LoRA updates still incurs substantial communication overheads. To reduce communication cost, FFA-LoRA (Sun et al., 2024) proposes a strategy of freezing the randomly initialized matrix $\mathbf{A}$ and only training the zero-initialized matrix $\mathbf{B}$. However, this method inevitably leads to model performance degradation due to the restricted parameter update space. Then, RoLoRA (Chen et al., 2025) introduces an alternating optimization strategy, where matrix $\mathbf{B}$ is updated and aggregated in odd communication rounds and matrix $\mathbf{A}$ is updated and aggregated in even rounds. This method emphasizes the importance of updating and aggregating the matrix $\mathbf{A}$, which mitigates the severe degradation of model performance. However, it only achieves the communication efficiency comparable to that of FFA-LoRA. To further reduce communication overhead, Fed-SB (Singhal et al., 2025) inserts a trainable square matrix $\mathbf{R}$ between the matrices $\mathbf{B}$, $\mathbf{A}$, then freezes the matrices $\mathbf{B}$, $\mathbf{A}$ and only trains and aggregates matrix $\mathbf{R}$. However, this method requires a larger rank than that of standard LoRA and relies on a computationally expensive initialization strategy.

Another line of work reduces communication overhead by compressing LoRA updates. FLASC (Kuo et al., 2026) introduces TopK sparsification into federated LoRA and transmits only LoRA updates with large magnitudes, but aggressive compression may degrade model performance. EcoLoRA (Liu et al., 2025) further combines round-robin segment sharing with adaptive TopK so that only sparse updates from a subset of LoRA modules are transmitted in each round. FedSRD (Yan et al., 2026) exploits the two-factor structure of LoRA to design an importance-aware sparsification strategy and generates sparse downlink updates through reconstruction and decomposition at the server. However, since different clients retain different update positions, their aggregated update may be

Table 1: Comparison of existing methods with ProLoRA. Communication cost is measured in 32-bit parameter equivalents, including binary masks. $d = (m+n)r$ is the LoRA parameter count, q is the compression rate, $c > 1$ is the rank scaling factor, and N_s is the number of segments. For simplicity, we assume $m = n$.

Method	Uplink	Downlink	Total	Trainable factors
FedIT	d	d	$2d$	$\mathbf{A}$, $\mathbf{B}$
FFA-LoRA	$d/2$	$d/2$	d	$\mathbf{B}$ only
RoLoRA	$d/2$	$d/2$	d	$\mathbf{A}$ / $\mathbf{B}$ (alternating)
Fed-SB	$c^2rd/(2m)$	$c^2rd/(2m)$	c^2rd/m	$\mathbf{R}$ only
EcoLoRA	$qd + d/(32N_s)$	$qd + d/32$	$2qd + (N_s+1)d/(32N_s)$	$\mathbf{A}$, $\mathbf{B}$
FedSRD	$qd + d/32$	$qd/2 + d/64$	$3qd/2 + 3d/64$	$\mathbf{A}$, $\mathbf{B}$
ProLoRA (Ours)	qd	qd	$2qd$	$\mathbf{A}$, $\mathbf{B}$

substantially denser than each individual update. Therefore, efficient downlink communication often requires additional sparsification, reconstruction, or decomposition at the server, which may introduce additional compression errors or computational overhead. In summary, most existing solutions face the dilemma that model performance degrades as communication overhead decreases. We highlight the advantages of ProLoRA in Table 1.

3 Preliminaries and Problem Formulation

In this paper, $\|\cdot\|$ denotes the Frobenius norm for a matrix or the ℓ_2 norm for a vector, with squared norms summed over factor blocks, and $[N]$ denotes the integer set $\{1, \ldots, N\}$. Due to space limitations, some preliminaries that are not directly essential to understanding the main text are deferred to Appendix A, including error feedback (EF) and random projection (RP).

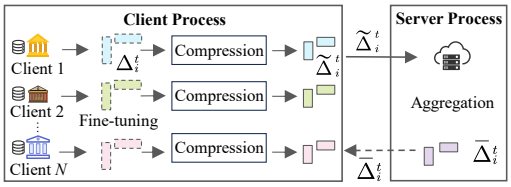


Figure 1: System model of ProLoRA.

3.1 Federated LoRA Fine-tuning and Biased Compression

In an FL scenario consisting of N clients and a server, federated LoRA fine-tuning aims to solve the following optimization problem:

$$\min_{\mathbf{B},\mathbf{A}} F(\mathbf{B},\mathbf{A}) := \frac{1}{N}\sum_{i=1}^{N} F_i(\mathbf{B},\mathbf{A}), \quad F_i(\mathbf{B},\mathbf{A}) := \mathbb{E}_{\xi_i\sim\mathcal{D}_i}[\mathcal{L}(\mathbf{W}_0 + \mathbf{B}\mathbf{A};\xi_i)], \tag{1}$$

where $\mathbf{W}_0 \in \mathbb{R}^{m\times n}$ denotes the frozen base model weights, $\mathbf{B} \in \mathbb{R}^{m\times r}$ denotes the zero-initialized LoRA matrix, $\mathbf{A} \in \mathbb{R}^{r\times n}$ represents the randomly initialized LoRA matrix, N denotes the total number of clients and ξ_i denotes a mini-batch randomly sampled from the local dataset $\mathcal{D}_i$ of client i. $\mathcal{L}$, F_i and F respectively denote the sample loss function, local loss function of client i and global loss function. Let $\boldsymbol{\theta} = (\mathbf{B},\mathbf{A}) \in \Phi = \mathbb{R}^{m\times r} \times \mathbb{R}^{r\times n}$, Eq. (1) aligns with the traditional FL optimization problem, allowing it to be solved via FedAvg. In FedAvg, the server and clients exchange the full LoRA module $(\mathbf{B},\mathbf{A})$ in each round. Thus, the communication costs of both uplink and downlink across T rounds with N clients are $\mathcal{O}(TN(m+n)r)$, which are substantial. To reduce communication overhead, many existing methods compress LoRA updates before transmission. A typical example is the TopK compressor $\mathcal{C}_k$, which retains the k entries with the largest absolute values and sets the rest to zero. We formalize the standard notion of biased compressors in Definition 1.

Definition 1 (Biased compressor (Alistarh et al., 2018; Stich et al., 2018; Li & Li, 2023))**.** *A (random) map $\mathcal{C} : \Phi \to \Phi$ is a biased compressor if there exists a constant $\omega_{\mathcal{C}} \in [0,1)$ such that $\mathbb{E}[\|\mathcal{C}(\boldsymbol{\theta}) - \boldsymbol{\theta}\|^2] \le \omega_{\mathcal{C}}^2\|\boldsymbol{\theta}\|^2$ for $\forall\boldsymbol{\theta} \in \Phi$.*

3.2 System Model

Figure 1 illustrates the system model of ProLoRA, which consists of a server and multiple clients. In round t, each client i first performs local training and computes the local LoRA update $\mathbf{\Delta}_i^t$. Then, each client i performs the compression mechanism using a compressor $\mathcal{C}$ that satisfies Definition 1 and uploads the compressed result $\widetilde{\mathbf{\Delta}}_i^t$ to the server. Finally, the server performs global aggregation and broadcasts the aggregated result $\mathbf{\Delta}^t$. In addition, the server and clients are semi-honest, which means that they honestly execute the protocols but may try to infer sensitive data.

3.3 Problem Definition

Existing communication-efficient federated LoRA methods usually reduce communication by restricting trainable parameters (Sun et al., 2024; Chen et al., 2025; Singhal et al., 2025) or compressing LoRA updates (Kuo et al., 2026; Liu et al., 2025; Yan et al., 2026). The former may limit the adaptation capacity of LoRA and degrade model performance, while the latter typically relies on sparsification. Since different clients retain different update positions, the aggregated global update often becomes much denser, resulting in high downlink communication cost or requiring server-side recompression with additional errors. Therefore, efficient bidirectional communication requires a compressor that supports direct aggregation in the compressed domain, i.e., $\frac{1}{N}\sum_{i=1}^{N}\mathcal{C}(\boldsymbol{\theta}_i) = \mathcal{C}(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\theta}_i)$.

However, biased compression still introduces information loss and therefore requires EF (Karimireddy et al., 2019; Li & Li, 2023) for compensation. Standard EF directly feeds back accumulated historical residuals, while LoRA consists of two coupled factors and the effect of a residual in one factor depends on the current state of the other factor. Therefore, under high compression and data heterogeneity, effective error compensation should further account for the coupled two-factor structure of LoRA.

4 Our Method: ProLoRA

4.1 Technical Overview

ROP for compact bidirectional communication. As discussed in Section 3.3, popular biased compressors such as TopK do not satisfy the linearity property. Motivated by the matrix structure and dimensional characteristics of LoRA, we design a linear compressor by exploiting the distributive property of matrix multiplication to compress the larger dimensions m and n. Reusing the same basis across rank channels reduces storage relative to a dense projection of vectorized factors.

RP provides a natural technical basis for this idea. Let $\mathbf{U} \in \mathbb{R}^{n\times s}$ be a random Gaussian matrix. Then the LoRA matrix $\mathbf{A} \in \mathbb{R}^{r\times n}$ can be projected as $\widetilde{\mathbf{A}} = \mathbf{A}\mathbf{U}$, where $\widetilde{\mathbf{A}} \in \mathbb{R}^{r\times s}$ and $s < n$. However, RP is typically used only for dimensionality reduction, whereas federated LoRA fine-tuning requires the projected representation to be restored to the original parameter space for continued training. A natural way to achieve this is to reconstruct $\mathbf{A}$ by $\widehat{\mathbf{A}} = \widetilde{\mathbf{A}}\mathbf{U}^{\dagger}$, where $\mathbf{U}^{\dagger}$ denotes the pseudoinverse of $\mathbf{U}$. Pseudoinverse computation typically relies on singular value decomposition (SVD), adding computational overhead (see Figure 4). Fortunately, for a column-orthonormal matrix $\mathbf{U}$, its pseudoinverse equals its transpose, i.e., $\mathbf{U}^{\dagger} = \mathbf{U}^{\top}$. Using this property, we design the ROP compressor $\mathcal{C}_p$ for LoRA.

Specifically, ROP is constructed via a projection–reconstruction process $\mathcal{C}_p = \mathcal{Q} \circ \mathcal{P}$. Let $\Omega = \mathbb{R}^{s\times r} \times \mathbb{R}^{r\times s}$, and let $\mathbf{U}_B \in \mathbb{R}^{m\times s}$ and $\mathbf{U}_A^{\top} \in \mathbb{R}^{n\times s}$ be independent random column-orthonormal matrices generated via Gaussian QR decomposition. The *projection operator* $\mathcal{P} : \Phi \to \Omega$, the *reconstruction operator* $\mathcal{Q} : \Omega \to \Phi$, and the ROP compressor $\mathcal{C}_p : \Phi \to \Phi$ are defined as follows:

$$\mathcal{P}(\boldsymbol{\theta}) = (\mathbf{U}_B^{\top}\mathbf{B}, \mathbf{A}\mathbf{U}_A^{\top}) = (\widetilde{\mathbf{B}}, \widetilde{\mathbf{A}}) = \widetilde{\boldsymbol{\theta}}, \tag{2}$$

$$\mathcal{Q}(\widetilde{\boldsymbol{\theta}}) = (\mathbf{U}_B\widetilde{\mathbf{B}}, \widetilde{\mathbf{A}}\mathbf{U}_A), \tag{3}$$

$$\mathcal{C}_p(\boldsymbol{\theta}) = \mathcal{Q}(\mathcal{P}(\boldsymbol{\theta})) = (\mathbf{U}_B(\mathbf{U}_B^{\top}\mathbf{B}), (\mathbf{A}\mathbf{U}_A^{\top})\mathbf{U}_A). \tag{4}$$

Proposition 1 establishes the linearity and compression error bound of ROP. Its linearity keeps the aggregated representation compact, enabling direct broadcast without server-side recompression. However, ROP remains a biased compressor and therefore requires EF to compensate for historical projection residuals.

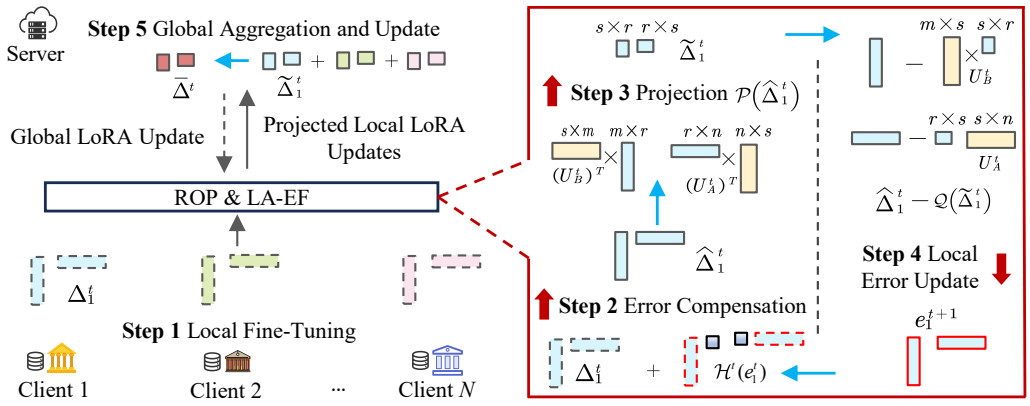


Figure 2: Overview of ProLoRA.

Proposition 1. *ROP is linear and satisfies Definition 1 with* $\omega_{\mathcal{C}}^2 = \max\{1 - s/m, 1 - s/n\}$.

LA-EF for LoRA-aware residual feedback. Although linear compression enables efficient bidirectional communication, it still introduces compression residuals. Standard EF directly feeds these historical residuals back into subsequent updates without distinguishing their different effects on the current LoRA module. For the global LoRA factors $(\mathbf{B}^t, \mathbf{A}^t)$ at round t, a residual in $\mathbf{B}$ affects the adapter through $\mathbf{e}_B\mathbf{A}^t$, while a residual in $\mathbf{A}$ acts through $\mathbf{B}^t\mathbf{e}_A$. These effects can be characterized by the Gram matrices of the corresponding partner factors:

$$\begin{aligned} \|\mathbf{e}_B\mathbf{A}^t\|^2 &= \mathrm{tr}\big(\mathbf{e}_B\mathbf{G}_B^t\mathbf{e}_B^\top\big), \quad \mathbf{G}_B^t = \mathbf{A}^t(\mathbf{A}^t)^\top, \\ \|\mathbf{B}^t\mathbf{e}_A\|^2 &= \mathrm{tr}\big(\mathbf{e}_A^\top\mathbf{G}_A^t\mathbf{e}_A\big), \quad \mathbf{G}_A^t = (\mathbf{B}^t)^\top\mathbf{B}^t. \end{aligned} \tag{5}$$

Therefore, historical residuals with the same parameter-space norm can have different effects on the current LoRA module. Under high compression and data heterogeneity, the uniform feedback of standard EF may hinder model performance recovery.

Motivated by this observation, we design LA-EF to regulate historical residual feedback using the global LoRA factors at the beginning of each round. For $b \in \{B, A\}$, we directly construct the residual filters and add only the filtered historical residual to the current update by Eq. 6.

$$\begin{aligned} \mathbf{H}_b^t &= \begin{cases} \left(\mathbf{I}_r + q\dfrac{r\mathbf{G}_b^t}{\mathrm{tr}(\mathbf{G}_b^t)}\right)^{-1}, & \mathrm{tr}(\mathbf{G}_b^t) > 0, \\ (1+q)^{-1}\mathbf{I}_r, & \mathrm{tr}(\mathbf{G}_b^t) = 0, \end{cases} \\ \mathcal{H}^t(\mathbf{e}_B, \mathbf{e}_A) &= (\mathbf{e}_B\mathbf{H}_B^t, \mathbf{H}_A^t\mathbf{e}_A), \qquad \widehat{\boldsymbol{\Delta}}_i^t = \boldsymbol{\Delta}_i^t + \mathcal{H}^t(\mathbf{e}_i^t). \end{aligned} \tag{6}$$

LA-EF regulates only the historical residual and leaves the current LoRA update $\boldsymbol{\Delta}_i^t$ unchanged. Residual directions with larger partner-factor energy are more strongly attenuated. Since the filters are computed entirely from the shared global LoRA factors, LA-EF introduces no additional communication overhead. Appendix B compares the residual-induced projection noise of standard EF and LA-EF with ROP.

4.2 Details of ProLoRA

The general procedure of ProLoRA is summarized in Algorithm 1. In round t, the server and all clients first generate the orthogonal projection basis $\mathbf{U}^t = (\mathbf{U}_B^t, \mathbf{U}_A^t)$ from the same pre-shared seed via QR decomposition, where $\mathbf{U}_B^t \in \mathbb{R}^{m\times s}$, $\mathbf{U}_A^t \in \mathbb{R}^{s\times n}$, and $s \ll \min(m, n)$. The remaining steps are described as follows.

1) Local Fine-Tuning. The server broadcasts the aggregated global LoRA update $\bar{\boldsymbol{\Delta}}^{t-1} \in \mathbb{R}^{s\times r} \times \mathbb{R}^{r\times s}$, whose projected size ensures downlink efficiency. Upon receiving $\bar{\boldsymbol{\Delta}}^{t-1}$, each client i reconstructs the global LoRA module $\boldsymbol{\theta}^t$ of round t according to Eq. (7) ($t > 1$).

$$\boldsymbol{\theta}^t = \boldsymbol{\theta}^{t-1} - \eta_g\mathcal{Q}^{t-1}(\bar{\boldsymbol{\Delta}}^{t-1}) = (\mathbf{B}^{t-1} - \eta_g\mathbf{U}_B^{t-1}\bar{\boldsymbol{\Delta}}_B^{t-1}, \mathbf{A}^{t-1} - \eta_g\bar{\boldsymbol{\Delta}}_A^{t-1}\mathbf{U}_A^{t-1}) = (\mathbf{B}^t, \mathbf{A}^t), \tag{7}$$

where $\mathbf{B}^t \in \mathbb{R}^{m\times r}$, $\mathbf{A}^t \in \mathbb{R}^{r\times n}$, and η_g denotes the global learning rate. Then, each client i performs local fine-tuning. In the e-th update step, the client i computes the stochastic gradient $\mathbf{g}_i^{t,e}$ by using a mini-batch $\xi_i^{t,e}$ randomly sampled from its local dataset $\mathcal{D}_i$ and local LoRA module $\boldsymbol{\theta}_i^{t,e}$ via Eq. (8).

$$\mathbf{g}_i^{t,e} = (\mathbf{g}_{B,i}^{t,e}, \mathbf{g}_{A,i}^{t,e}) = (\mathbf{F}_i^{t,e}(\mathbf{A}_i^{t,e})^\top, (\mathbf{B}_i^{t,e})^\top \mathbf{F}_i^{t,e}), \quad \mathbf{F}_i^{t,e} = \nabla\mathcal{L}((\mathbf{W}_0 + \mathbf{B}_i^{t,e}\mathbf{A}_i^{t,e}), \xi_i^{t,e}), \quad (8)$$

where $\mathbf{g}_{B,i}^{t,e} \in \mathbb{R}^{m\times r}$, $\mathbf{g}_{A,i}^{t,e} \in \mathbb{R}^{r\times n}$ and $\mathcal{L}$ is the sample loss function. After obtaining the stochastic gradient $\mathbf{g}_i^{t,e}$, the client i updates its local LoRA module $\boldsymbol{\theta}_i^{t,e+1}$ by Eq. (9).

$$\boldsymbol{\theta}_i^{t,e+1} = \boldsymbol{\theta}_i^{t,e} - \eta_l \mathbf{g}_i^{t,e} = (\mathbf{B}_i^{t,e} - \eta_l \mathbf{g}_{B,i}^{t,e}, \mathbf{A}_i^{t,e} - \eta_l \mathbf{g}_{A,i}^{t,e}) = (\mathbf{B}_i^{t,e+1}, \mathbf{A}_i^{t,e+1}), \quad (9)$$

where $\mathbf{B}_i^{t,e+1} \in \mathbb{R}^{m\times r}$, $\mathbf{A}_i^{t,e+1} \in \mathbb{R}^{r\times n}$, and η_l denotes the local learning rate. After E update steps of fine-tuning, each client i computes its local LoRA update $\boldsymbol{\Delta}_i^t$ according to Eq. (10).

$$\boldsymbol{\Delta}_i^t = \boldsymbol{\theta}^t - \boldsymbol{\theta}_i^{t,E+1} = (\mathbf{B}^t - \mathbf{B}_i^{t,E+1}, \mathbf{A}^t - \mathbf{A}_i^{t,E+1}) = (\boldsymbol{\Delta}_{B,i}^t, \boldsymbol{\Delta}_{A,i}^t). \quad (10)$$

2) Error Compensation. Each client i first computes the residual filters $\mathbf{H}_B^t$ and $\mathbf{H}_A^t$ from the global LoRA module $\boldsymbol{\theta}^t$ by Eq. (6). Each client i computes the error-compensated local LoRA update $\widehat{\boldsymbol{\Delta}}_i^t$ by combining the local LoRA update $\boldsymbol{\Delta}_i^t$ with the filtered local error module $\mathcal{H}^t(\mathbf{e}_i^t)$ via Eq. (11).

$$\widehat{\boldsymbol{\Delta}}_i^t = \boldsymbol{\Delta}_i^t + \mathcal{H}^t(\mathbf{e}_i^t) = (\boldsymbol{\Delta}_{B,i}^t + \mathbf{e}_{B,i}^t \mathbf{H}_B^t, \boldsymbol{\Delta}_{A,i}^t + \mathbf{H}_A^t \mathbf{e}_{A,i}^t) = (\widehat{\boldsymbol{\Delta}}_{B,i}^t, \widehat{\boldsymbol{\Delta}}_{A,i}^t). \quad (11)$$

3) Projection. Then, each client i projects $\widehat{\boldsymbol{\Delta}}_i^t$ using the projection operator $\mathcal{P}(\cdot)$ of the ROP compressor $\mathcal{C}_p$ by Eq. (12). Subsequently, each client sends its compressed result $\widetilde{\boldsymbol{\Delta}}_i^t$ to the server.

$$\widetilde{\boldsymbol{\Delta}}_i^t = \mathcal{P}^t(\widehat{\boldsymbol{\Delta}}_i^t) = ((\mathbf{U}_B^t)^\top \widehat{\boldsymbol{\Delta}}_{B,i}^t, \widehat{\boldsymbol{\Delta}}_{A,i}^t (\mathbf{U}_A^t)^\top) = (\widetilde{\boldsymbol{\Delta}}_{B,i}^t, \widetilde{\boldsymbol{\Delta}}_{A,i}^t). \quad (12)$$

4) Local Error Update. Furthermore, each client i must update its local error module $\mathbf{e}_i^{t+1}$ by using the reconstruction operator $\mathcal{Q}(\cdot)$ of the ROP compressor $\mathcal{C}_p$ according to Eq. (13). The updated error module $\mathbf{e}_i^{t+1}$ is stored locally and used in the next communication round.

$$\mathbf{e}_i^{t+1} = \widehat{\boldsymbol{\Delta}}_i^t - \mathcal{Q}^t(\widetilde{\boldsymbol{\Delta}}_i^t) = (\widehat{\boldsymbol{\Delta}}_{B,i}^t - \mathbf{U}_B^t \widetilde{\boldsymbol{\Delta}}_{B,i}^t, \widehat{\boldsymbol{\Delta}}_{A,i}^t - \widetilde{\boldsymbol{\Delta}}_{A,i}^t \mathbf{U}_A^t) = (\mathbf{e}_{B,i}^{t+1}, \mathbf{e}_{A,i}^{t+1}). \quad (13)$$

5) Global Aggregation and Update. The server computes the aggregated global LoRA update $\bar{\boldsymbol{\Delta}}^t$ by Eq. (14) and then updates the global LoRA module from $\boldsymbol{\theta}^t$ to $\boldsymbol{\theta}^{t+1}$.

$$\bar{\boldsymbol{\Delta}}^t = \frac{1}{N}\sum_{i=1}^N \widetilde{\boldsymbol{\Delta}}_i^t = (\frac{1}{N}\sum_{i=1}^N \widetilde{\boldsymbol{\Delta}}_{B,i}^t, \frac{1}{N}\sum_{i=1}^N \widetilde{\boldsymbol{\Delta}}_{A,i}^t) = (\bar{\boldsymbol{\Delta}}_B^t, \bar{\boldsymbol{\Delta}}_A^t). \quad (14)$$

5 Theoretical Analysis

We denote $\boldsymbol{\theta} = (\mathbf{B}, \mathbf{A}) \in \Phi = \mathbb{R}^{m\times r} \times \mathbb{R}^{r\times n}$ and rewrite $F(\mathbf{B}, \mathbf{A})$, $F_i(\mathbf{B}, \mathbf{A})$ as $F(\boldsymbol{\theta})$, $F_i(\boldsymbol{\theta})$, respectively. T, N and E represent the number of rounds, clients, and local update steps, respectively. Our analysis is based on the following assumptions in the non-convex setting.

Assumption 1 (Smoothness)**.** *The local loss function $F_i(\mathbf{X})$ on a fixed bounded convex region $\mathcal{K}$ covering all iterates and the factor-zero comparison points is L-smooth if there exists a positive constant L such that $\|\nabla F_i(\mathbf{X}) - \nabla F_i(\mathbf{Y})\| \le L\|\mathbf{X} - \mathbf{Y}\|$ for $\forall\, i \in [N]$, $\forall\, \mathbf{X}, \mathbf{Y} \in \mathcal{K}$.*

Assumption 2 (Bounded variance)**.** *The stochastic gradient $\mathbf{g}_i^{t,e}$ is unbiased such that $\mathbb{E}_{t,e}[\mathbf{g}_i^{t,e}] = \nabla F_i(\boldsymbol{\theta}_i^{t,e})$, and its variance v is bounded as $v = \mathbb{E}_{t,e}[\|\mathbf{g}_i^{t,e} - \nabla F_i(\boldsymbol{\theta}_i^{t,e})\|^2] \le \sigma^2$ for $\forall\, t \in [T]$, $\forall\, i \in [N]$, $\forall\, e \in [E]$, where $\mathbb{E}_{t,e}$ conditions on the pre-sampling history.*

Assumption 3 (Bounded gradient dissimilarity)**.** *The gradient dissimilarity d_g between global gradient $\nabla F(\boldsymbol{\theta}^t)$ and local gradient $\nabla F_i(\boldsymbol{\theta}^t)$ is bounded as $d_g = \frac{1}{N}\sum_{i=1}^N \|\nabla F_i(\boldsymbol{\theta}^t) - \nabla F(\boldsymbol{\theta}^t)\|^2 \le \delta^2$ for $\forall\, t \in [T]$, $\forall\, i \in [N]$.*

Algorithm 1 ProLoRA

Input: Communication rounds T, local update steps E, number N of clients, global learning rate η_g, local learning rate η_l, local LoRA rank r, projection dimension s, initial global LoRA module $\boldsymbol{\theta}^1 = (\mathbf{B}^1, \mathbf{A}^1) = (\mathbf{0}, \mathcal{N}(0, \sigma_g^2))$, initial local error modules $[\mathbf{e}_i^1]_{i=1}^N = (\mathbf{0}, \mathbf{0})$

for $t = 1, \ldots, T$ **do**
  Generate the orthogonal basis $\mathbf{U}^t$ from pre-shared random seeds via QR decomposition
  for client $i \in [N]$ **in parallel do**
    if $t > 1$ **then**
      Receive global LoRA update $\bar{\boldsymbol{\Delta}}^{t-1}$ from server
      Reconstruct global LoRA module $\boldsymbol{\theta}^t = \boldsymbol{\theta}^{t-1} - \eta_g \mathcal{Q}^{t-1}(\bar{\boldsymbol{\Delta}}^{t-1})$
    end if
    Set $\boldsymbol{\theta}_i^{t,1} = \boldsymbol{\theta}^t$
    for $e = 1, \ldots, E$ **do**
      Compute stochastic gradient $\mathbf{g}_i^{t,e}$ by Eq. (8)
      Update local LoRA module $\boldsymbol{\theta}_i^{t,e+1} = \boldsymbol{\theta}_i^{t,e} - \eta_l \mathbf{g}_i^{t,e}$
    end for
    Compute local LoRA update $\boldsymbol{\Delta}_i^t = \boldsymbol{\theta}^t - \boldsymbol{\theta}_i^{t,E+1}$
    Compute $\mathbf{H}_B^t, \mathbf{H}_A^t$ from $\boldsymbol{\theta}^t$ by Eq. (6)
    Perform error compensation $\widehat{\boldsymbol{\Delta}}_i^t = \boldsymbol{\Delta}_i^t + \mathcal{H}^t(\mathbf{e}_i^t)$
    Project error-compensated local LoRA update $\widetilde{\boldsymbol{\Delta}}_i^t = \mathcal{P}^t(\widehat{\boldsymbol{\Delta}}_i^t)$ and send $\widetilde{\boldsymbol{\Delta}}_i^t$ to server
    Update local error module $\mathbf{e}_i^{t+1} = \widehat{\boldsymbol{\Delta}}_i^t - \mathcal{Q}^t(\widetilde{\boldsymbol{\Delta}}_i^t)$
  end for
  Server performs global aggregation $\bar{\boldsymbol{\Delta}}^t = \frac{1}{N}\sum_{i=1}^N \widetilde{\boldsymbol{\Delta}}_i^t$
  Server performs global update $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \eta_g \mathcal{Q}^t(\bar{\boldsymbol{\Delta}}^t)$
end for

Assumption 1 specifies regional smoothness, which is specified in Appendix D.1 (Fang et al., 2025). Assumption 2 is standard in the convergence analysis of FL, while Assumption 3 is commonly adopted in FL to constrain data heterogeneity (Li & Li, 2023; Fang et al., 2025).

Theorem 1. *Let Assumptions 1 – 3 hold and* $\mathbf{e}_i^1 = \mathbf{0}$*, the iterates* $\{\boldsymbol{\theta}^t\}$ *generated by ProLoRA satisfy Eq.* (15) *on condition that* $\eta_l = \Theta(\frac{1}{E\sqrt{T}})$*,* $\eta_g = \Theta(\sqrt{EN})$ *and* $\eta_l \leq \frac{1}{8EL\cdot\max\{\sqrt{1+2/q}, \eta_g/q\}}$*.*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla F(\boldsymbol{\theta}^t)\|^2] = \mathcal{O}\left(\frac{\sqrt{q}\,\Delta_F + \sigma^2}{q^{3/2}\sqrt{TEN}} + \frac{\sigma^2 + E\delta^2}{qET} + \frac{(1-q)\sigma^2}{q^3 T}\right), \tag{15}$$

where $\Delta_F = F(\boldsymbol{\theta}^1) - F_\star$*,* $F_\star = \inf_{\boldsymbol{\theta}\in\mathcal{K}} F(\boldsymbol{\theta})$*, and* q*,* η_l *and* η_g *denote the compression ratio, local learning rate and global learning rate, respectively.*

Theorem 1 (proof in Appendix D.3) establishes a convergence upper bound. The upper bound contains no non-vanishing error term and vanishes as $T \to \infty$. A smaller q increases the upper bound. When $q = 1$, the residual terms vanish and ProLoRA recovers the convergence behavior of full-precision FedAvg.

6 Experiments

In this section, we comprehensively evaluate the performance of ProLoRA on natural language understanding (NLU) tasks. We compare ProLoRA with six baseline methods: FedIT (Zhang et al., 2024), FFA-LoRA (Sun et al., 2024), RoLoRA (Chen et al., 2025), Fed-SB (Singhal et al., 2025), EcoLoRA (Liu et al., 2025) and FedSRD (Yan et al., 2026).

Models and Datasets. For NLU tasks, we fine-tune and evaluate the RoBERTa-base (125M) (Liu et al., 2019) on five datasets from the GLUE (Wang et al., 2018) benchmark: MRPC, SST-2, QNLI, QQP and MNLI.

Settings. We use NVIDIA RTX A6000 or GeForce RTX 3090 GPUs for the main experiments. For the non-IID setting, we use the Dirichlet(α) distribution. Unless otherwise specified, all experimental

Table 2: Communication cost (MB) and performance (%) comparison on 5 GLUE tasks under the same number of communication rounds with RoBERTa-base.

Method	Comm. Cost (Up. / Down.)	MRPC	SST-2	QNLI	QQP	MNLI
FedIT	3375 / 3375	87.01±0.70	<u>93.76</u>±0.31	**91.61**±0.24	**86.92**±0.17	**84.76**±0.21
FFA-LoRA	1688 / 1688	86.41±0.56	92.58±0.13	88.91±0.96	84.51±0.28	80.50±0.61
RoLoRA	1688 / 1688	**88.45**±1.46	93.37±0.13	90.41±0.47	85.20±0.28	83.18±0.24
Fed-SB	<u>215</u> / <u>215</u>	82.87±2.17	90.25±0.76	85.67±0.82	80.59±0.91	54.59±10.22
EcoLoRA	440 / 2957	85.59±1.10	93.03±0.85	91.21±0.38	86.56±0.19	83.10±0.94
FedSRD	248 / 390	85.81±0.95	92.65±0.37	87.32±0.76	82.75±0.43	78.36±1.33
ProLoRA	**211 / 211**	<u>88.26</u>±1.05	**93.85**±0.46	<u>91.35</u>±0.59	<u>86.81</u>±0.32	<u>84.34</u>±0.49

Table 3: Performance (%) comparison on 5 GLUE tasks under the same reference communication budget (450 MB for total uplink and downlink) with RoBERTa-base.

Method	MRPC	SST-2	QNLI	QQP	MNLI	Avg.
FedIT	83.33±1.61	92.39±0.18	86.98±0.19	81.49±1.24	<u>77.67</u>±0.54	84.37
FFA-LoRA	<u>86.36</u>±1.23	92.01±0.24	84.00±1.80	80.91±1.12	73.89±2.18	83.43
RoLoRA	86.27±2.78	<u>92.51</u>±0.43	86.28±0.85	82.84±0.55	77.56±0.23	<u>85.09</u>
Fed-SB	82.84±2.14	90.25±0.70	85.70±0.87	80.59±0.90	54.57±10.21	78.79
EcoLoRA	84.40±1.67	91.74±1.39	<u>88.43</u>±0.28	<u>83.09</u>±0.09	76.17±3.44	84.76
FedSRD	86.19±0.93	92.47±0.37	86.81±0.96	82.25±0.42	76.84±1.42	84.91
ProLoRA	**88.32**±1.02	**93.85**±0.46	**91.31**±0.62	**86.81**±0.31	**84.35**±0.50	**88.93**

results are obtained with $N = 10$, $T = 300$, and $\alpha = 0.5$. For ProLoRA, each client uses the compression ratio $q = 1/16$. For EcoLoRA and FedSRD, we use their original adaptive compression parameters. We use rank $r = 28$ for Fed-SB on RoBERTa-base to approximately match the compression ratio of ProLoRA, while all other methods use rank $r = 8$. We run all experiments with three random seeds and report the mean and standard deviation. We select the local learning rate η_l from $\{1e-3, 5e-4, 1e-4, 5e-5, 2e-5\}$ and the global learning rate η_g from $\{1, 1.5, 2\}$ for all methods. In all tables, the best and second-best results are highlighted in **bold** and <u>underlined</u>, respectively. More detailed hyperparameter settings and information about the datasets are provided in Appendix F.

6.1 Results under Equal Communication Rounds

Table 2 compares the communication cost and model performance on five GLUE tasks under the same number of communication rounds, where LoRA parameters are encoded with 32 bits and TopK indices are counted using binary masks following FLASC (Kuo et al., 2026). ProLoRA reduces the total communication cost by 93.75%, compared with 50% for FFA-LoRA and RoLoRA, 49.66% for EcoLoRA, 90.54% for FedSRD, and 93.62% for Fed-SB. Meanwhile, ProLoRA maintains performance comparable to FedIT with an average increase of 0.11%, while the baselines exhibit performance degradation on most tasks, particularly on MNLI.

6.2 Results under Equal Communication Budget

To further illustrate the advantage of ProLoRA, we consider a scenario with a limited total communication budget. The budget is set equal to the total communication costs required by FedIT over 20 rounds. Under this budget constraint, we compare the performance of all methods. Table 3 shows that ProLoRA achieves the best performance on all tasks, improving the average performance by up to 10.14%. Its low per-round communication cost enables more training rounds under the same budget, leading to clear advantages over the baselines.

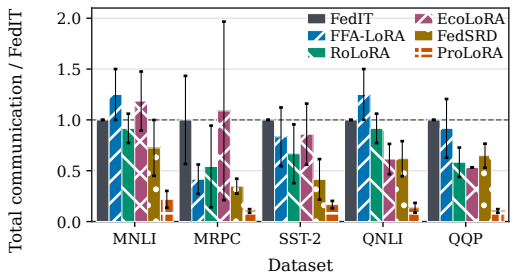


Figure 3: Communication cost under the corresponding target accuracy, normalized by FedIT.

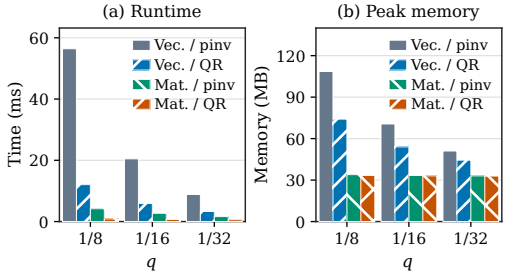


Figure 4: Runtime and peak GPU memory under different compression ratios q.

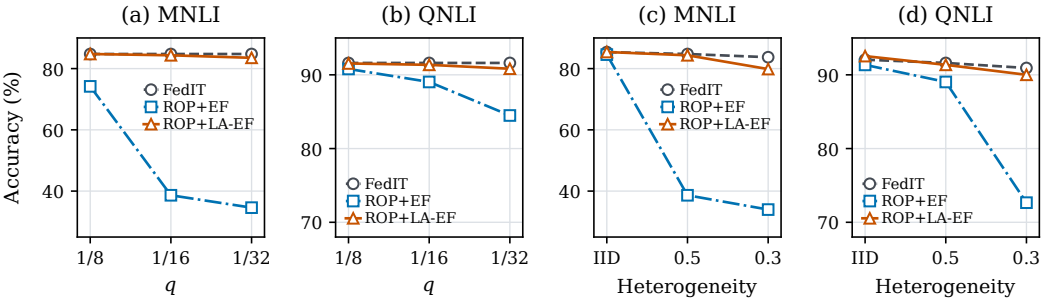


Figure 5: LA-EF ablation under different compression ratios and data heterogeneity.

6.3 Results under Equal Target Accuracy

Figure 3 compares the normalized communication cost required to reach the target accuracy on five tasks. ProLoRA achieves the lowest cost, reducing total communication by up to 80.44% compared with the best baseline. Fed-SB is omitted because it fails to reach the targets across all seeds.

6.4 Ablation Study

Ablation on the matrix design and orthogonality of ROP. Figure 4 shows that ROP achieves an 8.5× speedup over vector orthogonal projection and a 3.9× speedup over matrix non-orthogonal projection. The matrix design also reduces peak GPU memory from 54.03 MB to 33.06 MB, validating the efficiency of matrix projection and orthogonal reconstruction.

Ablation on LA-EF. Figure 5 shows that LA-EF consistently improves ROP over standard EF under different compression ratios and data heterogeneity. ProLoRA remains within 1.25% of FedIT for $q \in \{1/8, 1/16, 1/32\}$, while reducing total communication by up to 96.9%.

Further Experiments. Complete LA-EF recovery curves and results on larger models are reported in Appendices E.1 and E.2. Computation and memory overheads are discussed in Appendix E.3. We also study stronger data heterogeneity, more clients and larger LoRA ranks in Appendices E.4, E.5 and E.6, respectively.

7 Conclusion

In this paper, we propose ProLoRA, a communication-efficient federated LoRA fine-tuning framework that combines ROP with LA-EF to enable efficient bidirectional communication while preserving model performance. We provide a non-convex convergence analysis accounting for time-varying LoRA geometry and historical residual filtering, showing the same $\mathcal{O}(1/\sqrt{T})$ convergence order as full-precision FedAvg. Extensive experiments on multiple datasets and LLMs show that ProLoRA reduces total communication costs by up to 96.9% without sacrificing model performance.

AI USE STATEMENT

Large Language Models (LLMs) were used solely to assist with the writing and language polishing of this manuscript. Specifically, we used LLMs to improve grammar, sentence structure, readability, conciseness, and the overall flow of the text, as well as to rephrase sentences where necessary while preserving their original technical meaning. All AI-assisted edits were carefully reviewed and revised by the authors before being incorporated into the manuscript.

LLMs were not used for research ideation, methodology or algorithm design, theoretical derivations or proofs, code development, experimental design or execution, data generation, result analysis, or scientific interpretation. All research ideas, technical methods, theoretical results, experimental results, and conclusions presented in this paper were developed, verified, and approved by the authors. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work follows the ICLR Code of Ethics. The study does not involve human subjects, animal experimentation, or the collection of new private or sensitive data. All datasets and benchmarks used in our experiments are publicly available and are used in accordance with their corresponding usage guidelines. No personally identifiable information is used or collected in this work. We are not aware of additional ethical concerns beyond those generally associated with federated learning and large language models.

REPRODUCIBILITY STATEMENT

Our implementation code is publicly available through the anonymous repository at `https://anonymous.4open.science/r/ProLoRA-4F1D/`, which contains the code and configurations used to reproduce the main experiments in this paper. The complete ProLoRA algorithm and main theoretical results are provided in the main text, with the full convergence proof included in the appendix. Detailed hyperparameters, datasets, and model configurations are provided in Appendix F, while additional experiments, ablation studies, and computation and memory analyses are reported in Appendix E. Together with the anonymous repository, these materials support reproduction of the main theoretical and empirical results presented in this work.

REFERENCES

Dan Alistarh, Torsten Hoefler, Mikael Johansson, Nikola Konstantinov, Sarit Khirirat, and Cédric Renggli. The convergence of sparsified gradient methods. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5973–5983, 2018.

Shuangyi Chen, Yuanxin Guo, Yue Ju, Hardik Dalal, Zhongwen Zhu, and Ashish J Khisti. Robust federated finetuning of LLMs via alternating optimization of LoRA. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*, 2019.

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.

Yuze Han, Xiang Li, Shiyun Lin, and Zhihua Zhang. A random projection approach to personalized federated learning: enhancing communication efficiency, robustness, and fairness. *Journal of Machine Learning Research*, 25(380):1–88, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

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.

Sai Praneeth Karimireddy, Quentin Rebjock, Sebastian Stich, and Martin Jaggi. Error feedback fixes signsgd and other gradient compression schemes. In *Proceedings of International Conference on Machine Learning (ICML)*, pp. 3252–3261, 2019.

Kevin Kuo, Arian Raje, Kousik Rajesh, and Virginia Smith. FLASC: Federated LoRA with sparse communication. In *Proceedings of the ACM Conference on AI and Agentic Systems (CAIS)*, pp. 990–1008. ACM, 2026.

Shiwei Li, Xiandi Luo, Haozhao Wang, Xing Tang, Shijie Xu, Weihong Luo, Yuhua Li, Xiuqiang He, and Ruixuan Li. The panaceas for improving low-rank decomposition in communication-efficient federated learning. *arXiv preprint arXiv:2505.23176*, 2025.

Xiaoyun Li and Ping Li. Analysis of error feedback in federated non-convex optimization with biased compression: Fast convergence and partial participation. In *Proceedings of International Conference on Machine Learning (ICML)*, pp. 19638–19688, 2023.

Han Liu, Ruoyao Wen, Srijith Nair, Jia Liu, Wenjing Lou, Chongjie Zhang, William Yeoh, Yevgeniy Vorobeychik, and Ning Zhang. EcoLoRA: Communication-efficient federated fine-tuning of large language models. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 20732–20746, 2025.

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. *arXiv preprint arXiv:1907.11692*, 2019.

Yuren Mao, Yuhang Ge, Yijiang Fan, Wenyi Xu, Yu Mi, Zhonghao Hu, and Yunjun Gao. A survey on lora of large language models. *Frontiers of Computer Science*, 19(7):197605, 2025.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of Artificial Intelligence and Statistics (AISTATS)*, pp. 1273–1282, 2017.

Mohammad Hasan Narimani and Mostafa Tavassolipour. Fedrp: A communication-efficient approach for differentially private federated learning using random projection. *arXiv preprint arXiv:2509.10041*, 2025.

Raghav Singhal, Kaustubh Ponkshe, Rohit Vartak, Lav R Varshney, and Praneeth Vepakomma. FedSB: A silver bullet for extreme communication efficiency and performance in (private) federated lora fine-tuning. *arXiv preprint arXiv:2502.15436*, 2025.

Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified sgd with memory. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pp. 4452–4463, 2018.

Youbang Sun, Zitao Li, Yaliang Li, and Bolin Ding. Improving LoRA in privacy-preserving federated learning. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2024.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

Herbert Woisetschläger, Alexander Erben, Shiqiang Wang, Ruben Mayer, and Hans-Arno Jacobsen. A survey on efficient federated learning methods for foundation model training. In *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 8317–8325, 2024.

Yebo Wu, Chunlin Tian, Jingguang Li, He Sun, Kahou Tam, Zhanting Zhou, Haicheng Liao, Zhijiang Guo, Li Li, and Chengzhong Xu. A survey on federated fine-tuning of large language models. *arXiv preprint arXiv:2503.12016*, 2025.

Guochen Yan, Luyuan Xie, Qingni Shen, Yuejian Fang, and Zhonghai Wu. FedSRD: Sparsify-reconstruct-decompose for communication-efficient federated large language models fine-tuning. In *Proceedings of the ACM Web Conference (WWW)*, pp. 5087–5098. ACM, 2026.

Menglin Yang, Jialin Chen, Yifei Zhang, Jiahong Liu, Jiasheng Zhang, Qiyao Ma, Harshit Verma, Qianru Zhang, Min Zhou, Irwin King, et al. Low-rank adaptation for foundation models: A comprehensive review. *arXiv preprint arXiv:2501.00365*, 2024.

Rui Ye, Wenhao Wang, Jingyi Chai, Dihan Li, Zexi Li, Yinda Xu, Yaxin Du, Yanfeng Wang, and Siheng Chen. Openfedllm: Training large language models on decentralized private data via federated learning. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 6137–6147, 2024.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*, 2023.

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 *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6915–6919, 2024.

Appendix

A Preliminaries **13**

A.1 Error Feedback 13

A.2 Random Projection 13

B Residual-Noise Mechanism **14**

C Proof of Propositions **14**

C.1 Proof of Proposition 1 14

D Proof of Convergence Results **15**

D.1 Explanation of Assumption 1 15

D.2 Lemmas and Their Proofs 15

D.3 Proof of Theorem 1 21

E More experiment results **24**

E.1 Complete LA-EF Ablation Results 24

E.2 Results on Llama-2-7B and Gemma-7B 25

E.3 Computation and Memory Overhead 25

E.4 Effect of the degree of non-IID Data 27

E.5 Effect of the Number of Clients 27

E.6 Effect of the Rank of LoRA 28

F Implementation Details **28**

F.1 Details on Hyperparameters 28

F.2 Details on Datasets and Benchmarks 28

A Preliminaries

A.1 Error Feedback

Directly applying biased compression in FL may lead to divergence or severe performance degradation. EF (Karimireddy et al., 2019; Li & Li, 2023) is a standard technique to mitigate this issue. In EF, each client i maintains a local error term $\mathbf{e}_i^t$ that accumulates past compression errors. Given a local model update $\mathbf{x}_i^t$ and a biased compressor $\mathcal{C}$, client i first forms the error-compensated update $\widehat{\mathbf{x}}_i^t = \mathbf{x}_i^t + \mathbf{e}_i^t$, then compresses it as $\widetilde{\mathbf{x}}_i^t = \mathcal{C}(\widehat{\mathbf{x}}_i^t)$, and finally updates the local error by $\mathbf{e}_i^{t+1} = \widehat{\mathbf{x}}_i^t - \widetilde{\mathbf{x}}_i^t$. In this way, EF compensates compression errors accumulated in previous rounds and improves performance.

A.2 Random Projection

RP (Han et al., 2024; Narimani & Tavassolipour, 2025) is a dimensionality reduction technique that maps high-dimensional data to a lower-dimensional space using a random matrix. Let $\mathbf{x} \in \mathbb{R}^n$ be a data vector and $\mathbf{P} \in \mathbb{R}^{n \times k}$ be a random Gaussian matrix, where $k < n$. The projected result $\widetilde{\mathbf{x}} \in \mathbb{R}^k$ is computed as $\widetilde{\mathbf{x}} = \mathbf{P}^\top \mathbf{x}$. RP is widely used for dimensionality reduction, data compression, and efficient approximation in high-dimensional settings.

B Residual-Noise Mechanism

Let $J_t(\mathbf{X}) = \mathbf{X}_B\mathbf{A}^t + \mathbf{B}^t\mathbf{X}_A$ denote the first-order adapter change. For $m = n$ and $q = s/m = s/n$, the centered projection-noise energy of historical residuals satisfies

$$\mathbb{E}_{\mathcal{C}_p^t}\left[\left\|J_t\left((\mathcal{C}_p^t - q\mathcal{I})\mathbf{e}\right)\right\|^2\right] = q(1-q)\left(\|\mathbf{e}_B\mathbf{A}^t\|^2 + \|\mathbf{B}^t\mathbf{e}_A\|^2\right), \tag{16}$$

where $\mathcal{I}$ is the identity operator and the expectation is over fresh independent ROP bases.

Figure 6 compares residual-induced projection noise with and without LA-EF filtering on MNLI. LA-EF reduces this noise throughout training.

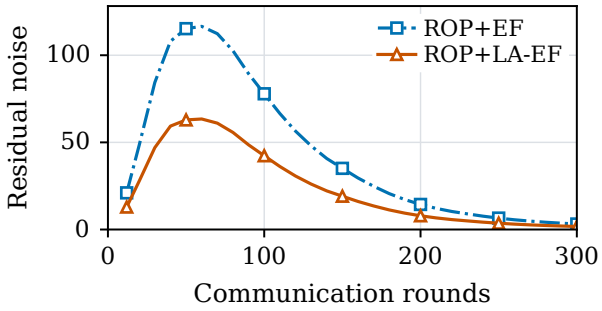


Figure 6: Residual-induced projection noise on MNLI with standard EF and LA-EF.

Let $\mu_b^t = \mathrm{tr}(\mathbf{G}_b^t)/r$ for $b \in \{B, A\}$. For $\mu_B^t > 0$, the B-side filter admits the proximal interpretation

$$\mathbf{e}_B\mathbf{H}_B^t = \arg\min_{\mathbf{Y}}\left\{\tfrac{1}{2}\|\mathbf{Y} - \mathbf{e}_B\|^2 + \tfrac{q}{2\mu_B^t}\|\mathbf{Y}\mathbf{A}^t\|^2\right\}. \tag{17}$$

Its first-order condition is $\mathbf{Y}(\mathbf{I}_r + q\mathbf{G}_B^t/\mu_B^t) = \mathbf{e}_B$. The A-side case is symmetric. Thus, LA-EF balances residual retention against its effect on the current adapter.

Let λ_j and $\mathbf{v}_j$ be the eigenvalues and eigenvectors of $\mathbf{G}_B^t$. Since $\mathbf{H}_B^t$ shares these eigenvectors, we have

$$\|\mathbf{e}_B\mathbf{H}_B^t\mathbf{A}^t\|^2 = \sum_{j=1}^{r}\frac{\lambda_j\|\mathbf{e}_B\mathbf{v}_j\|^2}{(1 + q\lambda_j/\mu_B^t)^2} \le \sum_{j=1}^{r}\lambda_j\|\mathbf{e}_B\mathbf{v}_j\|^2 = \|\mathbf{e}_B\mathbf{A}^t\|^2. \tag{18}$$

Together with the symmetric A-side inequality, this shows that LA-EF cannot increase the conditional noise in Eq. (16). If $\mu_b^t = 0$, the corresponding partner is zero and its noise term is zero.

To obtain Eq. (16), let $\mathbf{P}_B = \mathbf{U}_B\mathbf{U}_B^\top$ and $\mathbf{P}_A = \mathbf{U}_A^\top\mathbf{U}_A$. Haar isotropy gives $\mathbb{E}(\mathbf{P}_b - q\mathbf{I}) = \mathbf{0}$ and $\mathbb{E}[(\mathbf{P}_b - q\mathbf{I})^2] = q(1-q)\mathbf{I}$. Applying these identities to $(\mathbf{P}_B - q\mathbf{I})\mathbf{e}_B\mathbf{A}$ and $\mathbf{B}\mathbf{e}_A(\mathbf{P}_A - q\mathbf{I})$ yields the two squared-norm terms. Independence of the A/B bases makes the expected cross term zero.

C Proof of Propositions

C.1 Proof of Proposition 1

Proof. We first prove that $\mathcal{C}_p$ satisfies Definition 1. Recall that $\mathcal{C}_p(\boldsymbol{\theta}) = (\mathbf{U}_B\mathbf{U}_B^\top\mathbf{B}, \mathbf{A}\mathbf{U}_A^\top\mathbf{U}_A)$. Let $\mathbf{P}_B = \mathbf{U}_B\mathbf{U}_B^\top \in \mathbb{R}^{m\times m}$ and $\mathbf{P}_A = \mathbf{U}_A^\top\mathbf{U}_A \in \mathbb{R}^{n\times n}$ denote orthogonal projectors with rank s, we have Eq. (19).

$$\begin{aligned}\mathbb{E}[\|\mathcal{C}_p(\boldsymbol{\theta}) - \boldsymbol{\theta}\|^2] &= \mathbb{E}[\|(\mathbf{P}_B\mathbf{B} - \mathbf{B}, \mathbf{A}\mathbf{P}_A - \mathbf{A})\|^2] \\ &= \mathbb{E}[\|(\mathbf{I} - \mathbf{P}_B)\mathbf{B}\|^2] + \mathbb{E}[\|\mathbf{A}(\mathbf{I} - \mathbf{P}_A)\|^2].\end{aligned} \tag{19}$$

Since $\mathbf{P}_B$ and $\mathbf{P}_A$ are orthogonal projectors, we have $(\mathbf{I}-\mathbf{P}_B)^2 = \mathbf{I}-\mathbf{P}_B$ and $(\mathbf{I}-\mathbf{P}_A)^2 = \mathbf{I}-\mathbf{P}_A$. Moreover, because $\mathbf{U}_B$ and $\mathbf{U}_A^\top$ are Haar-distributed column-orthonormal matrices, isotropy implies

$\mathbb{E}[\mathbf{P}_B] = \frac{s}{m}\mathbf{I}_m$ and $\mathbb{E}[\mathbf{P}_A] = \frac{s}{n}\mathbf{I}_n$. Therefore, we have Eqs. 20 and (21).

$$\mathbb{E}[\|(\mathbf{I} - \mathbf{P}_B)\mathbf{B}\|^2] = \mathrm{tr}(\mathbf{B}^\top \mathbb{E}[\mathbf{I} - \mathbf{P}_B]\mathbf{B}) = (1 - \frac{s}{m})\|\mathbf{B}\|^2, \tag{20}$$

$$\mathbb{E}[\|\mathbf{A}(\mathbf{I} - \mathbf{P}_A)\|^2] = \mathrm{tr}(\mathbf{A}\mathbb{E}[\mathbf{I} - \mathbf{P}_A]\mathbf{A}^\top) = (1 - \frac{s}{n})\|\mathbf{A}\|^2. \tag{21}$$

Substituting Eq. (20) and Eq. (21) into Eq. (19), we obtain Eq. (22).

$$\begin{aligned}\mathbb{E}[\|\mathcal{C}_p(\boldsymbol{\theta}) - \boldsymbol{\theta}\|^2] &= (1 - \frac{s}{m})\|\mathbf{B}\|^2 + (1 - \frac{s}{n})\|\mathbf{A}\|^2 \\ &\leq \omega_{\mathcal{C}}^2(\|\mathbf{B}\|^2 + \|\mathbf{A}\|^2) = \omega_{\mathcal{C}}^2\|\boldsymbol{\theta}\|^2,\end{aligned} \tag{22}$$

where $\omega_{\mathcal{C}}^2 = \max\{1 - \frac{s}{m}, 1 - \frac{s}{n}\}$. Hence, $\mathcal{C}_p$ satisfies Definition 1. Next, we prove linearity. For any $\{\boldsymbol{\theta}_i = (\mathbf{B}_i, \mathbf{A}_i)\}_{i=1}^N$, we have Eq. (23).

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^N \mathcal{C}_p(\boldsymbol{\theta}_i) &= (\frac{1}{N}\sum_{i=1}^N \mathbf{U}_B\mathbf{U}_B^\top\mathbf{B}_i, \frac{1}{N}\sum_{i=1}^N \mathbf{A}_i\mathbf{U}_A^\top\mathbf{U}_A) \\ &= (\mathbf{U}_B\mathbf{U}_B^\top(\frac{1}{N}\sum_{i=1}^N \mathbf{B}_i), (\frac{1}{N}\sum_{i=1}^N \mathbf{A}_i)\mathbf{U}_A^\top\mathbf{U}_A) \\ &= \mathcal{C}_p(\frac{1}{N}\sum_{i=1}^N \boldsymbol{\theta}_i).\end{aligned} \tag{23}$$

The same distributive calculation gives $\mathcal{C}_p(a\mathbf{X} + b\mathbf{Y}) = a\mathcal{C}_p(\mathbf{X}) + b\mathcal{C}_p(\mathbf{Y})$ for arbitrary $a, b \in \mathbb{R}$ and $\mathbf{X}, \mathbf{Y} \in \Phi$, proving linearity. □

D Proof of Convergence Results

D.1 Explanation of Assumption 1

The local and global iterates, including their endpoints, lie almost surely in a fixed bounded convex region $\mathcal{K} \subseteq \Phi$. For each global iterate, $\mathcal{K}$ also contains the comparison points obtained by setting either factor of any one LoRA layer to zero while leaving all other factors unchanged. Each local objective F_i is differentiable on an open neighborhood of $\mathcal{K}$, and there exists $L > 0$ such that

$$\|\nabla F_i(\mathbf{X}) - \nabla F_i(\mathbf{Y})\| \leq L\|\mathbf{X} - \mathbf{Y}\|, \qquad i \in [N], \quad \mathbf{X}, \mathbf{Y} \in \mathcal{K}. \tag{24}$$

Convexity of $\mathcal{K}$ allows us to apply the descent lemma to each global update. Regional smoothness and boundedness of $\mathcal{K}$ also ensure $F_\star > -\infty$. Since each LoRA layer enters the objective through $\mathbf{BA}$, setting $\mathbf{A}$ to zero also makes its $\mathbf{B}$-gradient zero, and conversely. Smoothness then bounds each factor gradient by its partner norm, as shown in Eq. (50). This bound controls the changing Gram normalization in Eq. (57), even when a factor approaches zero.

The theorem assumes that the iterates remain in $\mathcal{K}$. It does not require a convex objective or introduce a projection step. We use no separate bounded-gradient assumption or Lipschitz assumption on the normalized Gram map. The analysis follows the local SGD updates and fixed LA-EF filter in Algorithm 1.

D.2 Lemmas and Their Proofs

Lemma 1 (ROP moments and residual filters)**.** *Let $\mathcal{C}_p^t$ use fresh shared Gaussian-QR bases with retention ratio q on both factor sides, independent of local sampling and the round-start history, with*

$$(\mathbf{U}_B^t)^\top\mathbf{U}_B^t = \mathbf{I}_s, \qquad \mathbf{U}_A^t(\mathbf{U}_A^t)^\top = \mathbf{I}_s. \tag{25}$$

For $b \in \{B, A\}$, write $\mathbf{S}_b^t = r\mathbf{G}_b^t/\,\mathrm{tr}(\mathbf{G}_b^t)$ when $\mathrm{tr}(\mathbf{G}_b^t) > 0$, and $\mathbf{S}_b^t = \mathbf{I}_r$ otherwise. Then $\mathbf{H}_b^t = (\mathbf{I}_r + q\mathbf{S}_b^t)^{-1}$. For every block vector $\mathbf{V}$ fixed before the fresh projection,

$$\mathbb{E}_p[\mathcal{C}_p^t\mathbf{V}] = q\mathbf{V}, \qquad \mathbb{E}_p[\|\mathcal{C}_p^t\mathbf{V}\|^2] = q\|\mathbf{V}\|^2, \qquad \mathbb{E}_p[\|(\mathcal{I} - \mathcal{C}_p^t)\mathbf{V}\|^2] = (1 - q)\|\mathbf{V}\|^2, \tag{26}$$

$$\frac{1}{1 + qr}\mathbf{I}_r \preceq \mathbf{H}_b^t \preceq \mathbf{I}_r, \qquad \|\mathcal{H}^t\mathbf{V}\| \leq \|\mathbf{V}\|. \tag{27}$$

Here $\mathbb{E}_p$ averages over the fresh projection conditional on all pre-projection quantities, and $\mathcal{I}$ is the identity operator.

Proof. For a rank-s Haar projector $\mathbf{P}$ in ambient dimension d_0, rotational invariance makes $\mathbb{E}\mathbf{P}$ a scalar multiple of $\mathbf{I}_{d_0}$. Taking traces gives $\mathbb{E}\mathbf{P} = (s/d_0)\mathbf{I}_{d_0}$. Since $\mathbf{P}^\top = \mathbf{P}^2 = \mathbf{P}$, this proves the three blockwise moment identities. The matrix $\mathbf{S}_b^t$ is positive semidefinite with trace r, including the zero-Gram case. Its eigenvalues lie in $[0, r]$, which gives (27). □

Lemma 2 ((Li & Li, 2023))**.** *Let Assumptions 1 – 3 hold. Let $\mathscr{F}_t$ denote the round-start history, write $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathscr{F}_t]$, and put $\mathbf{g}^t = \nabla F(\boldsymbol{\theta}^t)$. Clients use independent local sampling streams conditional on $\mathscr{F}_t$. If $\eta_l \le \frac{1}{8EL}$, then the average local drift satisfies Eq.* (28) *for any round $t \in [T]$ and local update step $e \in [E+1]$.*

$$\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_t[\|\boldsymbol{\theta}_i^{t,e} - \boldsymbol{\theta}^t\|^2] \le 5\eta_l^2 E(\sigma^2 + 6E\delta^2) + 30\eta_l^2 E^2\|\nabla F(\boldsymbol{\theta}^t)\|^2. \tag{28}$$

Proof. Fix i, t and denote $d_{i,h} = \mathbb{E}_t[\|\boldsymbol{\theta}_i^{t,h+1} - \boldsymbol{\theta}^t\|^2]$ for $0 \le h \le E$. Let $\mathscr{F}_{t,e}$ contain $\mathscr{F}_t$ and all clients' sampling history before local step e, so $\mathbb{E}_{t,e}[\cdot] = \mathbb{E}[\cdot \mid \mathscr{F}_{t,e}]$. The sampling errors $\boldsymbol{\varepsilon}_i^{t,e} = \mathbf{g}_i^{t,e} - \nabla F_i(\boldsymbol{\theta}_i^{t,e})$ are martingale differences, so $\mathbb{E}_t\|\sum_{a=1}^{h}\boldsymbol{\varepsilon}_i^{t,a}\|^2 \le h\sigma^2$. Splitting the local update into its gradient and noise parts gives

$$\begin{aligned} d_{i,h} &\le 2\eta_l^2 h\sum_{a=1}^{h}\mathbb{E}_t[\|\nabla F_i(\boldsymbol{\theta}_i^{t,a})\|^2] + 2\eta_l^2 h\sigma^2 \\ &\le 4\eta_l^2 h^2\|\nabla F_i(\boldsymbol{\theta}^t)\|^2 + 4\eta_l^2 hL^2\sum_{a=1}^{h} d_{i,a-1} + 2\eta_l^2 h\sigma^2. \end{aligned} \tag{29}$$

The first inequality uses $\|\mathbf{x}+\mathbf{y}\|^2 \le 2\|\mathbf{x}\|^2 + 2\|\mathbf{y}\|^2$ and Jensen's inequality. The second uses Assumption 1. Let $D_i = \max_{0\le h\le E} d_{i,h}$. Since $4L^2\eta_l^2E^2 \le 1/16$,

$$D_i \le \frac{64}{15}\eta_l^2E^2\|\nabla F_i(\boldsymbol{\theta}^t)\|^2 + \frac{32}{15}\eta_l^2E\sigma^2. \tag{30}$$

Averaging over clients and applying Assumption 3,

$$\frac{1}{N}\sum_i\|\nabla F_i(\boldsymbol{\theta}^t)\|^2 = \|\mathbf{g}^t\|^2 + \frac{1}{N}\sum_i\|\nabla F_i(\boldsymbol{\theta}^t) - \mathbf{g}^t\|^2 \le \|\mathbf{g}^t\|^2 + \delta^2. \tag{31}$$

Combining these inequalities and using $64/15 \le 30$, $32/15 \le 5$ proves (28).

□

Lemma 3. *Denote $\boldsymbol{\Delta}^t = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{\Delta}_i^t$. Let Assumptions 1 – 3 hold, $\mathbb{E}_t[\|\boldsymbol{\Delta}^t\|^2]$ is bounded by Eq.* (32) *for $\forall\, t \in [T]$, provided $\eta_l \le 1/(8EL)$.*

$$\begin{aligned} \mathbb{E}_t[\|\boldsymbol{\Delta}^t\|^2] &\le (2\eta_l^2E^2 + 120\eta_l^4E^4L^2)\mathbb{E}_t[\|\nabla F(\boldsymbol{\theta}^t)\|^2] \\ &\quad + 4\frac{\eta_l^2E}{N}\sigma^2 + 20\eta_l^4E^3L^2(\sigma^2 + 6E\delta^2). \end{aligned} \tag{32}$$

Define the average sampling error by

$$\boldsymbol{\nu}^t = \frac{\eta_l}{N}\sum_{i=1}^{N}\sum_{e=1}^{E}[\mathbf{g}_i^{t,e} - \nabla F_i(\boldsymbol{\theta}_i^{t,e})]. \tag{33}$$

Then

$$\mathbb{E}_t[\boldsymbol{\nu}^t] = \mathbf{0}, \qquad \mathbb{E}_t[\|\boldsymbol{\nu}^t\|^2] \le \frac{\eta_l^2E}{N}\sigma^2. \tag{34}$$

Proof. We have the following formula by the definition of $\mathbf{\Delta}^t$.

$$\begin{aligned}\mathbb{E}_t[\|\mathbf{\Delta}^t\|^2] =&\mathbb{E}_t[\|\frac{1}{N}\eta_l\sum_{i=1}^{N}\sum_{e=1}^{E}\mathbf{g}_i^{t,e}-\eta_l E\nabla F(\boldsymbol{\theta}^t)+\eta_l E\nabla F(\boldsymbol{\theta}^t)\|^2]\\ \leq&2\eta_l^2\mathbb{E}_t[\|\frac{1}{N}\sum_{i=1}^{N}\sum_{e=1}^{E}\mathbf{g}_i^{t,e}-E\nabla F(\boldsymbol{\theta}^t)\|^2]+2\eta_l^2E^2\mathbb{E}_t[\|\nabla F(\boldsymbol{\theta}^t)\|^2]\\ \leq&2\eta_l^2\mathbb{E}_t[\|\frac{1}{N}\sum_{i=1}^{N}\sum_{e=1}^{E}\mathbf{g}_i^{t,e}-\frac{E}{N}\sum_{i=1}^{N}\nabla F_i(\boldsymbol{\theta}^t)\|^2]+2\eta_l^2E^2\mathbb{E}_t[\|\nabla F(\boldsymbol{\theta}^t)\|^2]\\ \leq&\frac{2\eta_l^2}{N^2}\mathbb{E}_t[\|\sum_{i=1}^{N}\sum_{e=1}^{E}(\mathbf{g}_i^{t,e}-\nabla F_i(\boldsymbol{\theta}_i^{t,e})+\nabla F_i(\boldsymbol{\theta}_i^{t,e})-\nabla F_i(\boldsymbol{\theta}^t))\|^2]\\ &+2\eta_l^2E^2\mathbb{E}_t[\|\nabla F(\boldsymbol{\theta}^t)\|^2].\end{aligned}$$

The squared-norm term $\mathcal{T}_1$ in the last inequality can be bounded as

$$\begin{aligned}\mathcal{T}_1=&\mathbb{E}_t[\|\sum_{i=1}^{N}\sum_{e=1}^{E}(\mathbf{g}_i^{t,e}-\nabla F_i(\boldsymbol{\theta}_i^{t,e})+\nabla F_i(\boldsymbol{\theta}_i^{t,e})-\nabla F_i(\boldsymbol{\theta}^t))\|^2]\\ \leq&2\mathbb{E}_t[\|\sum_{i=1}^{N}\sum_{e=1}^{E}(\mathbf{g}_i^{t,e}-\nabla F_i(\boldsymbol{\theta}_i^{t,e}))\|^2]+2\mathbb{E}_t[\|\sum_{i=1}^{N}\sum_{e=1}^{E}(\nabla F_i(\boldsymbol{\theta}_i^{t,e})-\nabla F_i(\boldsymbol{\theta}^t))\|^2]\\ \overset{(a)}{\leq}&2NE\sigma^2+2NE\sum_{i=1}^{N}\sum_{e=1}^{E}\mathbb{E}_t[\|(\nabla F_i(\boldsymbol{\theta}_i^{t,e})-\nabla F_i(\boldsymbol{\theta}^t))\|^2]\\ \overset{(b)}{\leq}&2NE\sigma^2+2NEL^2\sum_{i=1}^{N}\sum_{e=1}^{E}\mathbb{E}_t[\|\boldsymbol{\theta}_i^{t,e}-\boldsymbol{\theta}^t\|^2]\\ \overset{(c)}{\leq}&2NE\sigma^2+10\eta_l^2N^2E^3L^2(\sigma^2+6E\delta^2)+60\eta_l^2N^2E^4L^2\mathbb{E}_t[\|\nabla F(\boldsymbol{\theta}^t)\|^2],\end{aligned}$$

where (a) uses Assumption 2, conditional zero mean over local steps, and independent local sampling streams across clients conditional on $\mathscr{F}_t$. These properties also give Eq. (34) without requiring independence between the A- and B-side gradients. The inequality (b) uses Assumption 1, and (c) follows from Lemma 2. Therefore, we have

$$\mathbb{E}_t[\|\mathbf{\Delta}^t\|^2]\leq(2\eta_l^2E^2+120\eta_l^4E^4L^2)\mathbb{E}_t[\|\nabla F(\boldsymbol{\theta}^t)\|^2]+4\frac{\eta_l^2E}{N}\sigma^2+20\eta_l^4E^3L^2(\sigma^2+6E\delta^2).$$

□

Lemma 4. *Let Assumptions 1 – 3 hold,* $\mathbb{E}_t[\langle\nabla F(\boldsymbol{\theta}^t),-\mathbf{\Delta}^t+\eta_lE\nabla F(\boldsymbol{\theta}^t)\rangle]$ *satisfies Eq.* (35) *on condition that* $\eta_l\leq\frac{1}{8EL}$.

$$\mathbb{E}_t[\langle\nabla F(\boldsymbol{\theta}^t),-\mathbf{\Delta}^t+\eta_lE\nabla F(\boldsymbol{\theta}^t)\rangle]\leq\frac{3\eta_lE}{4}\mathbb{E}_t[\|\nabla F(\boldsymbol{\theta}^t)\|^2]+\frac{5\eta_l^3E^2L^2}{2}(\sigma^2+6E\delta^2).\tag{35}$$

Let $\widetilde{\sigma}^2=\sigma^2+6E\delta^2$ *and define the local-drift remainder by*

$$\mathbf{b}^t=\frac{\eta_l}{N}\sum_{i=1}^{N}\sum_{e=1}^{E}[\nabla F_i(\boldsymbol{\theta}_i^{t,e})-\nabla F_i(\boldsymbol{\theta}^t)].\tag{36}$$

It also satisfies

$$\mathbb{E}_t[\|\mathbf{b}^t\|^2]\leq5L^2\eta_l^4E^3\widetilde{\sigma}^2+30L^2\eta_l^4E^4\|\mathbf{g}^t\|^2.\tag{37}$$

Proof. By Jensen's inequality, Assumption 1, and Lemma 2,

$$\begin{aligned}\mathbb{E}_t[\|\mathbf{b}^t\|^2]\leq&\frac{\eta_l^2EL^2}{N}\sum_{i=1}^{N}\sum_{e=1}^{E}\mathbb{E}_t[\|\boldsymbol{\theta}_i^{t,e}-\boldsymbol{\theta}^t\|^2]\\ \leq&\eta_l^2E^2L^2[5\eta_l^2E\widetilde{\sigma}^2+30\eta_l^2E^2\|\mathbf{g}^t\|^2].\end{aligned}\tag{38}$$

This proves Eq. (37).

We have the following formula by the definition of $\langle \nabla F(\boldsymbol{\theta}^t), -\boldsymbol{\Delta}^t + \eta_l E \nabla F(\boldsymbol{\theta}^t)\rangle$.

$$
\begin{aligned}
&\mathbb{E}_t[\langle \nabla F(\boldsymbol{\theta}^t), -\boldsymbol{\Delta}^t + \eta_l E \nabla F(\boldsymbol{\theta}^t)\rangle] \\
&= \langle \nabla F(\boldsymbol{\theta}^t), \mathbb{E}_t[-\frac{1}{N}\eta_l \sum_{i=1}^{N}\sum_{e=1}^{E} \mathbf{g}_i^{t,e} + \eta_l E \nabla F(\boldsymbol{\theta}^t)]\rangle \\
&= \langle \sqrt{\eta_l}\nabla F(\boldsymbol{\theta}^t), \sqrt{\eta_l}\mathbb{E}_t[-\frac{1}{N}\sum_{i=1}^{N}\sum_{e=1}^{E} \nabla F_i(\boldsymbol{\theta}_i^{t,e}) + E\nabla F(\boldsymbol{\theta}^t)]\rangle \\
&\overset{(a)}{\leq} \frac{\eta_l}{2E}\mathbb{E}_t[\|\frac{1}{N}\sum_{i=1}^{N}\sum_{e=1}^{E}(\nabla F_i(\boldsymbol{\theta}_i^{t,e}) - \nabla F_i(\boldsymbol{\theta}^t))\|^2] + \frac{\eta_l E}{2}\mathbb{E}_t[\|\nabla F(\boldsymbol{\theta}^t)\|^2] \\
&\leq \frac{\eta_l}{2N}\sum_{i=1}^{N}\sum_{e=1}^{E}\mathbb{E}_t[\|\nabla F_i(\boldsymbol{\theta}_i^{t,e}) - \nabla F_i(\boldsymbol{\theta}^t)\|^2] + \frac{\eta_l E}{2}\mathbb{E}_t[\|\nabla F(\boldsymbol{\theta}^t)\|^2] \\
&\overset{(b)}{\leq} \frac{\eta_l L^2}{2N}\sum_{i=1}^{N}\sum_{e=1}^{E}\mathbb{E}_t[\|\boldsymbol{\theta}_i^{t,e} - \boldsymbol{\theta}^t\|^2] + \frac{\eta_l E}{2}\mathbb{E}_t[\|\nabla F(\boldsymbol{\theta}^t)\|^2] \\
&\overset{(c)}{\leq} \frac{\eta_l E L^2}{2}\left[5\eta_l^2 E(\sigma^2 + 6E\delta^2) + 30\eta_l^2 E^2 \mathbb{E}_t[\|\nabla F(\boldsymbol{\theta}^t)\|^2]\right] + \frac{\eta_l E}{2}\mathbb{E}_t[\|\nabla F(\boldsymbol{\theta}^t)\|^2] \\
&\overset{(d)}{\leq} \frac{47\eta_l E}{64}\mathbb{E}_t[\|\nabla F(\boldsymbol{\theta}^t)\|^2] + \frac{5\eta_l^3 E^2 L^2}{2}(\sigma^2 + 6E\delta^2),
\end{aligned}
$$

where (a) is due to $\langle \mathbf{a}, \mathbf{b}\rangle \leq \frac{\rho}{2}\|\mathbf{a}\|^2 + \frac{1}{2\rho}\|\mathbf{b}\|^2$ for $\forall \rho \in \mathbb{R}$ and $\rho > 0$, (b) uses Assumption 1, (c) uses Lemma 2 and (d) is due to $\eta_l \leq \frac{1}{8EL}$. The proof is completed since $\frac{47}{64} < \frac{3}{4}$. □

Lemma 5. *Let* $\mathbf{e}^t = N^{-1}\sum_i \mathbf{e}_i^t$, *with* $\mathbf{e}_i^1 = \mathbf{0}$. *Define the auxiliary noise residual by*

$$
\mathbf{n}^1 = \mathbf{0}, \qquad \mathbf{n}^{t+1} = (\mathcal{I} - \mathcal{C}_p^t)(\boldsymbol{\nu}^t + \mathcal{H}^t \mathbf{n}^t). \tag{39}
$$

Let Assumptions 1 – 3 hold, $\mathbb{E}[\|\mathbf{e}^{t+1}\|^2]$ *is bounded via Eq.* (40)*, provided* $\eta_l \leq 1/(8EL)$.

$$
\begin{aligned}
\mathbb{E}[\|\mathbf{e}^{t+1}\|^2] \leq \frac{C_1}{q}\sum_{\tau=1}^{t}\left(1 - \frac{q}{2}\right)^{t-\tau}\mathbb{E}[\|\nabla F(\boldsymbol{\theta}^\tau)\|^2] \\
+ \frac{16\eta_l^2(1-q)E}{q^2 N}\sigma^2 + \frac{80\eta_l^4(1-q)E^3L^2}{q^2}(\sigma^2 + 6E\delta^2),
\end{aligned} \tag{40}
$$

where $C_1 = 4(1-q)\eta_l^2 E^2(60\eta_l^2 E^2 L^2 + 1)$ *and* $\mathbf{e}^{t+1} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{e}_i^{t+1}$. *Moreover,*

$$
\mathbb{E}[\|\mathbf{n}^t\|^2] \leq \frac{1-q}{q}\frac{\eta_l^2 E^2 \sigma^2}{NE}, \tag{41}
$$

$$
\sum_{t=1}^{T}\mathbb{E}[\|\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t\|^2] \leq \frac{4\eta_g^2}{q}\sum_{t=1}^{T}\mathbb{E}[\|\boldsymbol{\Delta}^t\|^2]. \tag{42}
$$

Proof. First suppose $0 < q < 1$. Condition on all pre-projection quantities and average over the fresh ROP. Young's inequality with parameter $q/[2(1-q)]$ yields

$$
\begin{aligned}
\mathbb{E}_p[\|\mathbf{e}^{t+1}\|^2] &= (1-q)\|\boldsymbol{\Delta}^t + \mathcal{H}^t\mathbf{e}^t\|^2 \\
&\leq \left(1 - \frac{q}{2}\right)\|\mathbf{e}^t\|^2 + \frac{(1-q)(2-q)}{q}\|\boldsymbol{\Delta}^t\|^2 \\
&\leq \left(1 - \frac{q}{2}\right)\|\mathbf{e}^t\|^2 + \frac{2(1-q)}{q}\|\boldsymbol{\Delta}^t\|^2.
\end{aligned} \tag{43}
$$

The filter contraction in Lemma 1 is used in the second line. Taking full expectations and unrolling from $\mathbf{e}^1 = \mathbf{0}$ gives

$$\mathbb{E}[\|\mathbf{e}^{t+1}\|^2] \le \frac{2(1-q)}{q} \sum_{\tau=1}^{t} \left(1 - \frac{q}{2}\right)^{t-\tau} \mathbb{E}[\|\mathbf{\Delta}^\tau\|^2]. \tag{44}$$

Substitute Lemma 3 and bound the geometric sum by $2/q$ to obtain (40). For the noise residual, average first over the projection and then over the current local sampling:

$$\begin{aligned} \mathbb{E}_t[\|\mathbf{n}^{t+1}\|^2] &= (1-q)\big(\mathbb{E}_t[\|\boldsymbol{\nu}^t\|^2] + \|\mathcal{H}^t\mathbf{n}^t\|^2\big) \\ &\le (1-q)\left(\frac{\eta_l^2 E^2 \sigma^2}{NE} + \|\mathbf{n}^t\|^2\right). \end{aligned} \tag{45}$$

The cross term vanishes because $\mathcal{H}^t\mathbf{n}^t$ is $\mathscr{F}_t$-measurable and $\mathbb{E}_t\boldsymbol{\nu}^t = \mathbf{0}$. A geometric sum from $\mathbf{n}^1 = \mathbf{0}$ proves (41). Next sum the middle line of (43). Since $\mathbf{e}^1 = \mathbf{0}$,

$$\begin{aligned} \sum_{t=1}^{T} \mathbb{E}[\|\mathbf{e}^t\|^2] &\le \sum_{t=1}^{T} \mathbb{E}[\|\mathbf{e}^{t+1}\|^2], \\ \sum_{t=1}^{T} \mathbb{E}[\|\mathbf{e}^{t+1}\|^2] &\le \frac{2(1-q)(2-q)}{q^2} \sum_{t=1}^{T} \mathbb{E}[\|\mathbf{\Delta}^t\|^2]. \end{aligned} \tag{46}$$

By the projector moments,

$$\begin{aligned} \sum_{t=1}^{T} \mathbb{E}[\|\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t\|^2] &= \frac{\eta_g^2 q}{1-q} \sum_{t=1}^{T} \mathbb{E}[\|\mathbf{e}^{t+1}\|^2] \\ &\le \frac{2(2-q)\eta_g^2}{q} \sum_{t=1}^{T} \mathbb{E}[\|\mathbf{\Delta}^t\|^2] \le \frac{4\eta_g^2}{q} \sum_{t=1}^{T} \mathbb{E}[\|\mathbf{\Delta}^t\|^2]. \end{aligned} \tag{47}$$

This proves (42). When $q = 1$, $\mathcal{C}_p^t = \mathcal{I}$, so $\mathbf{e}^t = \mathbf{n}^t = \mathbf{0}$ and the three conclusions follow directly without division by $1 - q$.

□

Lemma 6 (Normalized Gram matrices and gradient variation)**.** *Under Assumption 1, let $\mathbf{S}_b^t$ be defined in Lemma 1, and let $\|\cdot\|_{\mathrm{op}}$ denote the induced operator norm. Define*

$$\mathbf{M}_b^t = (\mathbf{I}_r + \mathbf{S}_b^t)^{-1}, \qquad \mathcal{M}^t(\mathbf{X}_B, \mathbf{X}_A) = (\mathbf{X}_B\mathbf{M}_B^t, \mathbf{M}_A^t\mathbf{X}_A), \tag{48}$$

$$\boldsymbol{\psi}^t = \mathcal{M}^t\mathbf{g}^t. \tag{49}$$

For every global iterate and each LoRA layer,

$$\|\nabla_{\mathbf{B}} F(\mathbf{B}, \mathbf{A})\| \le L\|\mathbf{A}\|, \qquad \|\nabla_{\mathbf{A}} F(\mathbf{B}, \mathbf{A})\| \le L\|\mathbf{B}\|, \tag{50}$$

$$\mathcal{M}^t[\mathcal{I} - (1-q)\mathcal{H}^t] = q\mathcal{H}^t, \qquad \|\mathcal{M}^t\|_{\mathrm{op}} \le 1, \tag{51}$$

$$\|(\mathcal{M}^{t+1} - \mathcal{M}^t)\mathbf{g}^t\| \le 4rL\|\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t\|, \tag{52}$$

$$\|\boldsymbol{\psi}^{t+1} - \boldsymbol{\psi}^t\| \le (1+4r)L\|\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t\|. \tag{53}$$

Proof. Setting this layer's $\mathbf{A}$ to zero makes the objective independent of its $\mathbf{B}$, so the corresponding $\mathbf{B}$-gradient is zero. Both points lie in $\mathcal{K}$ by Assumption 1. Applying (24) gives the first inequality in (50). Interchanging the two factors gives the second. Other layers remain unchanged in each comparison.

For any eigenvalue $x \geq 0$ of $\mathbf{S}_b^t$,

$$\frac{1}{1+x}\left(1-\frac{1-q}{1+qx}\right)=\frac{q}{1+qx}. \tag{54}$$

This proves Eq. (51), including the zero-Gram case. To bound the variation, write $\mathbf{M}(\mathbf{Z})=(\mathbf{I}_r+\mathbf{S}(\mathbf{Z}))^{-1}$, where $\mathbf{S}(\mathbf{Z})=r\mathbf{Z}\mathbf{Z}^\top/\|\mathbf{Z}\|^2$ for $\mathbf{Z}\neq\mathbf{0}$ and $\mathbf{S}(\mathbf{0})=\mathbf{I}_r$. If $\mathbf{Z},\mathbf{Z}'\neq\mathbf{0}$, set $\mathbf{X}=\mathbf{Z}/\|\mathbf{Z}\|$ and $\mathbf{Y}=\mathbf{Z}'/\|\mathbf{Z}'\|$. Then

$$\begin{aligned}\|\mathbf{X}-\mathbf{Y}\| &\leq \frac{2\|\mathbf{Z}'-\mathbf{Z}\|}{\|\mathbf{Z}\|},\\ \|\mathbf{S}(\mathbf{Z}')-\mathbf{S}(\mathbf{Z})\|_{\mathrm{op}} &\leq r(\|\mathbf{X}\|_{\mathrm{op}}+\|\mathbf{Y}\|_{\mathrm{op}})\|\mathbf{X}-\mathbf{Y}\| \leq \frac{4r\|\mathbf{Z}'-\mathbf{Z}\|}{\|\mathbf{Z}\|}.\end{aligned} \tag{55}$$

The first inequality follows by adding and subtracting $\mathbf{Z}'/\|\mathbf{Z}\|$. The second uses the difference of the two normalized Gramians. The inverse-matrix identity gives

$$\mathbf{M}(\mathbf{Z}')-\mathbf{M}(\mathbf{Z})=\mathbf{M}(\mathbf{Z}')\left[\mathbf{S}(\mathbf{Z})-\mathbf{S}(\mathbf{Z}')\right]\mathbf{M}(\mathbf{Z}). \tag{56}$$

Since $0\prec\mathbf{M}\preceq\mathbf{I}_r$,

$$\|\mathbf{M}(\mathbf{Z}')-\mathbf{M}(\mathbf{Z})\|_{\mathrm{op}} \leq \min\left\{1,\frac{4r\|\mathbf{Z}'-\mathbf{Z}\|}{\|\mathbf{Z}\|}\right\}. \tag{57}$$

If $\mathbf{Z}'=\mathbf{0}$ and $\mathbf{Z}\neq\mathbf{0}$, the bound by 1 proves the same inequality. If $\mathbf{Z}=\mathbf{0}$, the corresponding old factor gradient is zero by (50), so that block contributes zero. For the B side use $\mathbf{Z}=\mathbf{A}^t$, and for the A side use $\mathbf{Z}=(\mathbf{B}^t)^\top$. Multiplying (57) by (50) cancels the partner norm in the denominator:

$$\begin{aligned}\|\nabla_{\mathbf{B}}F(\boldsymbol{\theta}^t)(\mathbf{M}_B^{t+1}-\mathbf{M}_B^t)\| &\leq 4rL\|\mathbf{A}^{t+1}-\mathbf{A}^t\|,\\ \|(\mathbf{M}_A^{t+1}-\mathbf{M}_A^t)\nabla_{\mathbf{A}}F(\boldsymbol{\theta}^t)\| &\leq 4rL\|\mathbf{B}^{t+1}-\mathbf{B}^t\|.\end{aligned} \tag{58}$$

Summing the squared block bounds proves (52). Finally,

$$\boldsymbol{\psi}^{t+1}-\boldsymbol{\psi}^t=\mathcal{M}^{t+1}(\mathbf{g}^{t+1}-\mathbf{g}^t)+(\mathcal{M}^{t+1}-\mathcal{M}^t)\mathbf{g}^t. \tag{59}$$

Assumption 1, $\|\mathcal{M}^{t+1}\|_{\mathrm{op}}\leq 1$ and Eq. (52) imply Eq. (53), completing the proof.

□

Lemma 7 (Historical noise identity)**.** *Under Assumptions 1 – 3, let $\mathbf{n}^t$ and $\boldsymbol{\psi}^t$ be defined in Lemmas 5 and 6, respectively. Then*

$$q\sum_{t=1}^{T}\mathbb{E}[\langle\mathbf{g}^t,\mathcal{H}^t\mathbf{n}^t\rangle]=-\mathbb{E}[\langle\boldsymbol{\psi}^T,\mathbf{n}^{T+1}\rangle]+\sum_{t=2}^{T}\mathbb{E}[\langle\boldsymbol{\psi}^t-\boldsymbol{\psi}^{t-1},\mathbf{n}^t\rangle], \tag{60}$$

Proof. Define $\boldsymbol{\zeta}^t=\mathbf{n}^{t+1}-(1-q)\mathcal{H}^t\mathbf{n}^t$. Averaging (39) first over the fresh projection and then over local sampling gives $\mathbb{E}_t\boldsymbol{\zeta}^t=\mathbf{0}$. By (51) and self-adjointness,

$$q\langle\mathbf{g}^t,\mathcal{H}^t\mathbf{n}^t\rangle=\langle\boldsymbol{\psi}^t,\mathbf{n}^t-\mathbf{n}^{t+1}\rangle+\langle\boldsymbol{\psi}^t,\boldsymbol{\zeta}^t\rangle. \tag{61}$$

The last term has zero expectation because $\boldsymbol{\psi}^t$ is $\mathscr{F}_t$-measurable. Summation by parts and $\mathbf{n}^1=\mathbf{0}$ prove (60). □

D.3 Proof of Theorem 1

Proof. Denote $\mathbf{e}^t = N^{-1}\sum_{i=1}^N \mathbf{e}_i^t$ and $\mathbf{\Delta}^t = N^{-1}\sum_{i=1}^N \mathbf{\Delta}_i^t$. Since all clients use the same linear maps $\mathcal{P}^t$, $\mathcal{Q}^t$ and $\mathcal{H}^t$, the update rules in Algorithm 1 give

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \eta_g \mathcal{Q}^t\left(\frac{1}{N}\sum_i \mathcal{P}^t(\mathbf{\Delta}_i^t + \mathcal{H}^t\mathbf{e}_i^t)\right) = \boldsymbol{\theta}^t - \eta_g \mathcal{C}_p^t(\mathbf{\Delta}^t + \mathcal{H}^t\mathbf{e}^t),$$
$$\mathbf{e}^{t+1} = \frac{1}{N}\sum_i (\mathcal{I} - \mathcal{C}_p^t)(\mathbf{\Delta}_i^t + \mathcal{H}^t\mathbf{e}_i^t) = (\mathcal{I} - \mathcal{C}_p^t)(\mathbf{\Delta}^t + \mathcal{H}^t\mathbf{e}^t). \tag{62}$$

Thus, with $\widehat{\mathbf{\Delta}}^t = \mathbf{\Delta}^t + \mathcal{H}^t\mathbf{e}^t$,

$$\mathcal{C}_p^t\widehat{\mathbf{\Delta}}^t + \mathbf{e}^{t+1} + (\mathcal{I} - \mathcal{H}^t)\mathbf{e}^t = \mathbf{\Delta}^t + \mathbf{e}^t. \tag{63}$$

All non-LoRA parameters are fixed, and fixed LoRA scaling is absorbed into the loss. Norms and inner products sum over factor blocks. We use the same rank r and retention ratio q across layers, and the fresh shared bases in Lemma 1 are independent of the past and local sampling.

Set $\eta = E\eta_l$ and $\mathbf{g}^t = \nabla F(\boldsymbol{\theta}^t)$. By the local update rule and the definitions of $\mathbf{b}^t$ and $\boldsymbol{\nu}^t$ in Lemmas 4 and 3,

$$\mathbf{\Delta}^t = \eta\mathbf{g}^t + \mathbf{b}^t + \boldsymbol{\nu}^t. \tag{64}$$

Let $\mathbf{n}^t$ be the noise residual in Lemma 5, and put $\mathbf{s}^t = \mathbf{e}^t - \mathbf{n}^t$. Define

$$\mathbf{c}^t = \eta\mathbf{g}^t + \mathbf{b}^t + \mathcal{H}^t\mathbf{s}^t, \qquad \mathbf{w}^t = \boldsymbol{\nu}^t + \mathcal{H}^t\mathbf{n}^t. \tag{65}$$

Subtracting Eq. (39) from the residual update in Eq. (62) gives

$$\widehat{\mathbf{\Delta}}^t = \mathbf{c}^t + \mathbf{w}^t, \qquad \mathbf{s}^{t+1} = (\mathcal{I} - \mathcal{C}_p^t)\mathbf{c}^t. \tag{66}$$

Write $K_{\mathrm{LA}} = 1 + 2/q$ and $\widetilde{\sigma}^2 = \sigma^2 + 6E\delta^2$. The learning-rate condition in Theorem 1 is

$$0 < \eta_l \le \frac{1}{8EL\max\{\sqrt{K_{\mathrm{LA}}}, \eta_g/q\}}. \tag{67}$$

We consider the auxiliary sequence

$$\lambda = \frac{\eta_g q(2+q)}{4\eta}, \qquad \kappa = \frac{\eta_g q^2(1+2q)}{8\eta}, \qquad \mathcal{W}^t = F(\boldsymbol{\theta}^t) + \lambda\|\mathbf{s}^t\|^2. \tag{68}$$

According to Assumption 1, we have

$$F(\boldsymbol{\theta}^{t+1}) \le F(\boldsymbol{\theta}^t) + \langle \mathbf{g}^t, \boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t\rangle + \frac{L}{2}\|\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t\|^2. \tag{69}$$

Taking expectation over the current projection and local sampling, using Lemma 1 and $\|\mathbf{c} + \mathbf{w}\|^2 \le 2\|\mathbf{c}\|^2 + 2\|\mathbf{w}\|^2$, we obtain

$$\begin{aligned}\mathbb{E}_t[\mathcal{W}^{t+1}] - \mathcal{W}^t \le\ & -\eta_g q\mathbb{E}_t[\langle \mathbf{g}^t, \mathbf{c}^t\rangle] + [L\eta_g^2 q + \lambda(1-q)]\mathbb{E}_t[\|\mathbf{c}^t\|^2] - \lambda\|\mathbf{s}^t\|^2 \\ & + L\eta_g^2 q\mathbb{E}_t[\|\mathbf{w}^t\|^2] - q\eta_g\langle \mathbf{g}^t, \mathcal{H}^t\mathbf{n}^t\rangle.\end{aligned} \tag{70}$$

Here we use $\mathbb{E}_t\boldsymbol{\nu}^t = \mathbf{0}$. Since $\mathbf{c}^t - \eta\mathbf{g}^t = \mathbf{b}^t + \mathcal{H}^t\mathbf{s}^t$,

$$-\eta_g q\langle \mathbf{g}^t, \mathbf{c}^t\rangle = -\frac{\eta_g q}{2\eta}\left[\|\mathbf{c}^t\|^2 + \eta^2\|\mathbf{g}^t\|^2 - \|\mathbf{b}^t + \mathcal{H}^t\mathbf{s}^t\|^2\right], \tag{71}$$

$$\|\mathbf{b}^t + \mathcal{H}^t\mathbf{s}^t\|^2 \le K_{\mathrm{LA}}\|\mathbf{b}^t\|^2 + \left(1 + \frac{q}{2}\right)\|\mathcal{H}^t\mathbf{s}^t\|^2. \tag{72}$$

After substitution, the coefficient of $\mathbb{E}_t\|\mathbf{c}^t\|^2$ is

$$\frac{\eta_g q}{2\eta}\left[2L\eta_g\eta - \frac{q(1+q)}{2}\right] \le -\frac{\eta_g q^2(1+2q)}{8\eta} = -\kappa, \tag{73}$$

because (67) implies $L\eta_g\eta \le q/8$. The remaining signal-residual term is $\lambda(\|\mathcal{H}^t\mathbf{s}^t\|^2 - \|\mathbf{s}^t\|^2) \le 0$. Consequently,

$$\begin{aligned}\mathbb{E}_t[\mathcal{W}^{t+1}] - \mathcal{W}^t \le & -\frac{q\eta_g\eta}{2}\|\mathbf{g}^t\|^2 - \kappa\mathbb{E}_t[\|\mathbf{c}^t\|^2] + \frac{q\eta_g K_{\mathrm{LA}}}{2\eta}\mathbb{E}_t[\|\mathbf{b}^t\|^2] \\ & + L\eta_g^2 q\mathbb{E}_t[\|\mathbf{w}^t\|^2] - q\eta_g\langle\mathbf{g}^t, \mathcal{H}^t\mathbf{n}^t\rangle.\end{aligned} \tag{74}$$

By Lemma 4, the gradient contribution of the local-drift term is at most $15q\eta_g K_{\mathrm{LA}}L^2\eta^3\|\mathbf{g}^t\|^2$. Since $K_{\mathrm{LA}}L^2\eta^2 \le 1/64$, the remaining descent coefficient is at least $17q\eta_g\eta/64 \ge q\eta_g\eta/4$. Also,

$$\begin{aligned}\mathbb{E}[\|\mathbf{w}^t\|^2] &= \mathbb{E}[\|\boldsymbol{\nu}^t\|^2] + \mathbb{E}[\|\mathcal{H}^t\mathbf{n}^t\|^2] \\ &\le \frac{\eta^2\sigma^2}{NE} + \frac{1-q}{q}\frac{\eta^2\sigma^2}{NE} = \frac{\eta^2\sigma^2}{qNE}.\end{aligned} \tag{75}$$

The equality follows from round-start measurability and conditional zero mean, and the inequality uses Lemmas 3 and 5. Substituting these bounds into Eq. (74) and taking full expectations, we obtain

$$\begin{aligned}\mathbb{E}[\mathcal{W}^{t+1}] - \mathbb{E}[\mathcal{W}^t] \le & -\frac{q\eta_g\eta}{4}\mathbb{E}[\|\mathbf{g}^t\|^2] - \kappa\mathbb{E}[\|\mathbf{c}^t\|^2] + \frac{5}{2}q\eta_g K_{\mathrm{LA}}L^2\eta_l^3E^2\widetilde{\sigma}^2 \\ & + \frac{L\eta_g^2\eta^2\sigma^2}{NE} - q\eta_g\mathbb{E}[\langle\mathbf{g}^t, \mathcal{H}^t\mathbf{n}^t\rangle].\end{aligned} \tag{76}$$

Let $\mathcal{A}_T = \sum_{t=1}^T \mathbb{E}[\|\mathbf{g}^t\|^2], \mathcal{U}_T = \sum_{t=1}^T \mathbb{E}[\|\mathbf{c}^t\|^2]$ and $\Delta_F = F(\boldsymbol{\theta}^1) - F_\star$. Summing Eq. (76) over $t = 1, \dots, T$ and using $\mathcal{W}^1 = F(\boldsymbol{\theta}^1)$ and $\mathcal{W}^{T+1} \ge F_\star$, we obtain

$$\begin{aligned}\frac{q\eta_g\eta}{4}\mathcal{A}_T + \kappa\mathcal{U}_T \le & \Delta_F + \frac{5}{2}q\eta_g K_{\mathrm{LA}}L^2\frac{\eta^3 T}{E}\widetilde{\sigma}^2 + \frac{L\eta_g^2\eta^2\sigma^2 T}{NE} \\ & - q\eta_g\sum_{t=1}^T\mathbb{E}[\langle\mathbf{g}^t, \mathcal{H}^t\mathbf{n}^t\rangle].\end{aligned} \tag{77}$$

By Lemma 7, the last term in Eq. (77) consists of a terminal term and a sum of gradient variations. The terminal term is bounded by Young's inequality and Lemma 5 as

$$\begin{aligned}\eta_g|\mathbb{E}[\langle\boldsymbol{\psi}^T, \mathbf{n}^{T+1}\rangle]| &\le \frac{q\eta_g\eta}{16}\mathbb{E}[\|\mathbf{g}^T\|^2] + \frac{4\eta_g}{q\eta}\mathbb{E}[\|\mathbf{n}^{T+1}\|^2] \\ &\le \frac{q\eta_g\eta}{16}\mathcal{A}_T + \frac{4(1-q)\eta_g\eta}{q^2NE}\sigma^2.\end{aligned} \tag{78}$$

For the cumulative term, let $D_{\mathrm{LA}} = (1+4r)L\sqrt{1-q}$ and write $\mathcal{V}_T = \sum_{t=1}^T \mathbb{E}[\|\mathbf{w}^t\|^2] \le \eta^2\sigma^2 T/(qNE)$ by Eq. (75). The projector moments and Minkowski's inequality give

$$\begin{aligned}\sum_{t=1}^T\mathbb{E}[\|\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t\|^2] &= \eta_g^2 q\sum_{t=1}^T\mathbb{E}[\|\mathbf{c}^t + \mathbf{w}^t\|^2] \\ &\le \eta_g^2 q(\sqrt{\mathcal{U}_T} + \sqrt{\mathcal{V}_T})^2.\end{aligned} \tag{79}$$

Combining this inequality with Eqs. (53) and (41), Cauchy–Schwarz gives

$$\eta_g \left| \sum_{t=2}^{T} \mathbb{E}[\langle \boldsymbol{\psi}^t - \boldsymbol{\psi}^{t-1}, \mathbf{n}^t \rangle] \right| \leq D_{\mathrm{LA}} \eta_g^2 \eta \sigma \sqrt{\frac{T}{NE}} (\sqrt{\mathcal{U}_T} + \sqrt{\mathcal{V}_T}). \tag{80}$$

Young's inequality bounds the first term on the right-hand side as

$$D_{\mathrm{LA}} \eta_g^2 \eta \sigma \sqrt{\frac{T}{NE}} \sqrt{\mathcal{U}_T} \leq \kappa \mathcal{U}_T + \frac{2 D_{\mathrm{LA}}^2 \eta_g^3 \eta^3 \sigma^2 T}{q^2 (1+2q) NE}. \tag{81}$$

The bound on $\mathcal{V}_T$ gives $D_{\mathrm{LA}} \eta_g^2 \eta^2 \sigma^2 T / (\sqrt{q}\, NE)$ for the second term. Consequently,

$$\eta_g \left| \sum_{t=2}^{T} \mathbb{E}[\langle \boldsymbol{\psi}^t - \boldsymbol{\psi}^{t-1}, \mathbf{n}^t \rangle] \right| \leq \kappa \mathcal{U}_T + \frac{2 D_{\mathrm{LA}}^2 \eta_g^3 \eta^3 \sigma^2 T}{q^2 (1+2q) NE} + \frac{D_{\mathrm{LA}} \eta_g^2 \eta^2 \sigma^2 T}{\sqrt{q}\, NE}. \tag{82}$$

Substituting Eqs. (78) and (82) into Eq. (77) cancels $\kappa \mathcal{U}_T$. Since $1/4 - 1/16 \geq 1/8$, rearranging gives

$$\begin{aligned} \frac{q \eta_g \eta}{8} \mathcal{A}_T \leq \Delta_F + \frac{5}{2} q \eta_g K_{\mathrm{LA}} L^2 \frac{\eta^3 T}{E} \widetilde{\sigma}^2 + \frac{(L + D_{\mathrm{LA}}/\sqrt{q}) \eta_g^2 \eta^2 \sigma^2 T}{NE} \\ + \frac{4(1-q) \eta_g \eta \sigma^2}{q^2 NE} + \frac{2 D_{\mathrm{LA}}^2 \eta_g^3 \eta^3 \sigma^2 T}{q^2 (1+2q) NE}. \end{aligned} \tag{83}$$

Dividing by $q \eta_g \eta T / 8$ and substituting $\eta = E \eta_l$, we have

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\|\nabla F(\boldsymbol{\theta}^t)\|^2] \leq \frac{8 \Delta_F}{q \eta_g \eta_l E T} + \frac{8 (L + D_{\mathrm{LA}}/\sqrt{q}) \eta_g \eta_l}{qN} \sigma^2 \\ + 20 K_{\mathrm{LA}} L^2 E \eta_l^2 \widetilde{\sigma}^2 + \frac{32(1-q)}{q^3 NET} \sigma^2 + \frac{16 D_{\mathrm{LA}}^2 \eta_g^2 E \eta_l^2}{q^3 (1+2q) N} \sigma^2. \end{aligned} \tag{84}$$

When $q = 1$, both residuals are zero and $D_{\mathrm{LA}} = 0$. The same bound follows with the residual terms removed.

Set $\eta_l = c_l / (E\sqrt{T})$ and $\eta_g = c_g \sqrt{EN}$, where $c_l, c_g > 0$ are fixed constants satisfying Eq. (67). Substitution gives

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\|\nabla F(\boldsymbol{\theta}^t)\|^2] \leq \frac{8 \Delta_F}{q c_l c_g \sqrt{TEN}} + \frac{8 c_l c_g (L + D_{\mathrm{LA}}/\sqrt{q}) \sigma^2}{q \sqrt{TEN}} \\ + \frac{20 K_{\mathrm{LA}} L^2 c_l^2 \widetilde{\sigma}^2}{ET} + \left(\frac{32(1-q)}{q^3 NE} + \frac{16 D_{\mathrm{LA}}^2 c_l^2 c_g^2}{q^3 (1+2q)} \right) \frac{\sigma^2}{T}. \end{aligned} \tag{85}$$

Since $K_{\mathrm{LA}} \leq 3/q$, $D_{\mathrm{LA}} = (1+4r) L \sqrt{1-q}$, $NE \geq 1$ and $\widetilde{\sigma}^2 = \sigma^2 + 6E\delta^2$, Eq. (85) implies Eq. (15). The hidden constants depend only on L, r, c_l, c_g, not on q, T, E, N. Thus, ProLoRA achieves the stationarity rate $\mathcal{O}(T^{-1/2})$ for fixed E, N, q, r, completing the proof.

□

E MORE EXPERIMENT RESULTS

E.1 COMPLETE LA-EF ABLATION RESULTS

In Figure 7, we conduct ablation studies to evaluate the effect of LA-EF in ProLoRA. Figures 7(a)–(c) show the results on MNLI dataset. Figures 7(d)–(f) show the results on QNLI dataset. We observe that ProLoRA suffers from clear performance degradation with standard EF. In contrast, with LA-EF, ProLoRA maintains performance within 1.25% of full-precision FedIT across compression ratios from $q = 1/8$ to $q = 1/32$, demonstrating the effectiveness of combining the ROP compressor with LA-EF.

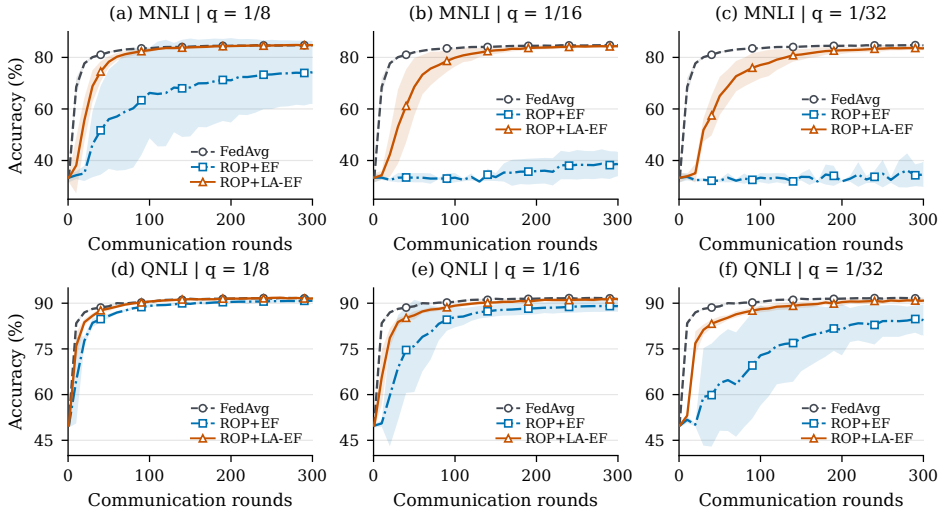


Figure 7: Recovery curves under different compression ratios.

Figure 8 further reports the complete recovery curves under IID data and Dirichlet $\alpha \in \{0.5, 0.3\}$. We observe that ROP+EF recovers well under IID data but degrades as data heterogeneity increases. LA-EF improves recovery under both non-IID settings. At $\alpha = 0.3$, it improves terminal accuracy from 33.97% to 79.84% on MNLI and from 72.66% to 90.62% on QNLI.

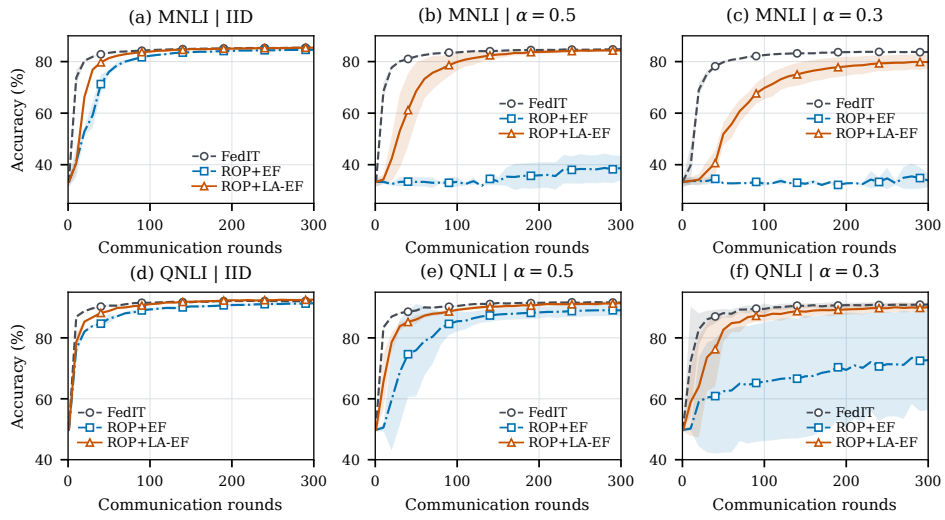


Figure 8: Recovery curves under different data partitions.

E.2 Results on Llama-2-7B and Gemma-7B

In the main manuscript, our experiments primarily focus on the RoBERTa-base model. To demonstrate the effectiveness of ProLoRA on more complex models, we conduct experiments on the Llama-2-7B and Gemma-7B models. Specifically, for the Llama-2-7B (Touvron et al., 2023) model, we fine-tune it on the Dolly-15K (Zhang et al., 2024) dataset and evaluate it on the MMLU (Hendrycks et al., 2020) and DROP (Dua et al., 2019) benchmarks. For the Gemma-7B (Team et al., 2024) model, we fine-tune it on the MetaMathQA (Yu et al., 2023) dataset and evaluate it on the GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021) benchmarks. All results are obtained under the settings of $N = 10$, $T = 600$ and $\alpha = 0.5$. we use compression ratio of $q = 1/16$. Except for Fed-SB, the local rank is set to $r = 16$ for all other methods. For Fed-SB, the local rank is $r = 91$. The experimental results are summarized in Table 4.

We observe that the total communication overhead also rises significantly as the model size increases. Specifically, FedIT's total communication overhead reaches 375 GB for Llama-2-7B and 246.1 GB for Gemma-7B. This further emphasizes the importance of addressing the communication bottleneck. From the results, ProLoRA restores the performance of FedIT, demonstrating its effectiveness on more complex models. In contrast, other methods exhibit performance degradation. Notably, ProLoRA reduces communication overhead by approximately 93.8%, significantly outperforming all baselines. These results further highlight the effectiveness and superiority of ProLoRA.

Table 4: Communication cost (GB) and score comparison on MMLU, DROP, GSM8K and MATH benchmarks under the same number of communication rounds with Llama-2-7B and Gemma-7B.

Method	Llama-2-7B				Gemma-7B			
	Comm. Cost (GB)		Score		Comm. Cost (GB)		Score	
	Up.	Down.	MMLU	DROP	Up.	Down.	GSM8K	MATH
FedIT	187.50	187.50	**43.63**$\pm$0.36	**31.32**$\pm$0.98	123.05	123.05	**55.23**$\pm$1.23	**11.41**$\pm$1.34
FFA-LoRA	93.75	93.75	42.73$\pm$0.26	30.35$\pm$1.21	61.53	61.53	52.98$\pm$0.75	10.55$\pm$0.45
RoLoRA	93.75	93.75	43.08$\pm$0.27	30.58$\pm$0.37	61.53	61.53	53.74$\pm$0.64	10.67$\pm$0.17
Fed-SB	11.85	11.85	40.32$\pm$0.14	28.07$\pm$0.62	8.12	8.12	50.78$\pm$0.33	8.54$\pm$0.59
EcoLoRA	28.76	170.63	43.13$\pm$0.11	30.83$\pm$0.31	20.46	115.28	54.84$\pm$0.25	10.12$\pm$0.71
FedSRD	16.88	21.56	42.13$\pm$1.41	29.37$\pm$1.69	9.84	14.15	51.36$\pm$1.39	9.85$\pm$1.61
ProLoRA	**11.72**	**11.72**	43.43$\pm$0.34	31.14$\pm$0.46	**7.69**	**7.69**	55.03$\pm$0.17	10.78$\pm$0.67

E.3 Computation and Memory Overhead

Although ProLoRA aims to reduce the communication cost of federated LoRA fine-tuning, it is still important to examine its computational overhead. We argue that a truly communication-efficient method should not introduce substantial additional computation to the system. Therefore, we measure the total runtime and peak GPU memory per communication round on RoBERTa-base and Llama-2-7B, and further examine the cost of each stage in ProLoRA.

We use one A6000 GPU with ten clients. RoBERTa-base uses a batch size of 32, 20 local steps and $r = 8$, while Llama-2-7B uses a batch size of 8, ten local steps and $r = 16$. Fed-SB uses ranks 28 and 91, respectively. Compression settings follow the main experiments. We average seven rounds after two warmup rounds, excluding initialization, evaluation and checkpoint saving.

Comparison with baselines. Table 5 reports the computation and memory costs of all methods. On RoBERTa-base, FFA-LoRA achieves the lowest runtime and peak GPU memory. ProLoRA increases the average round time from 19.17 s to 19.48 s compared with FedIT, an overhead of approximately 1.6%, while maintaining the same peak GPU memory of 2.48 GB. On Llama-2-7B, ProLoRA requires 140.86 s per round compared with 132.99 s for FedIT, an overhead of 5.92%, with the same peak GPU memory of 35.61 GB.

ProLoRA time breakdown. Figure 9 reports the additional time spent on ROP, LA-EF, aggregation and model update. ROP includes basis generation, projection and reconstruction, while LA-EF includes filter construction, filtering and residual update. Local training accounts for 98.53% and

Table 5: Average runtime per round and peak allocated GPU memory.

Method	RoBERTa-base		Llama-2-7B	
	Round time (s)	Peak GPU (GB)	Round time (s)	Peak GPU (GB)
FedIT	19.17$\pm$0.36	2.48	132.99$\pm$4.59	35.61
FFA-LoRA	18.83$\pm$0.28	2.20	133.08$\pm$4.57	31.58
RoLoRA	18.96$\pm$0.24	2.48	133.63$\pm$4.33	35.56
Fed-SB	19.26$\pm$0.43	2.22	130.53$\pm$4.60	31.78
EcoLoRA	19.59$\pm$0.46	2.48	137.32$\pm$4.56	35.61
FedSRD	20.98$\pm$0.46	2.48	567.14$\pm$5.73	35.61
ROP+EF	19.37$\pm$0.24	2.48	139.72$\pm$4.53	35.61
ProLoRA	19.48$\pm$0.45	2.48	140.86$\pm$4.59	35.61

93.30% of the round time on RoBERTa-base and Llama-2-7B, respectively, at $q = 1/16$. As q decreases from $1/8$ to $1/32$, the ROP cost decreases from 178.27 ms to 48.19 ms on RoBERTa-base and from 16.89 s to 3.14 s on Llama-2-7B.

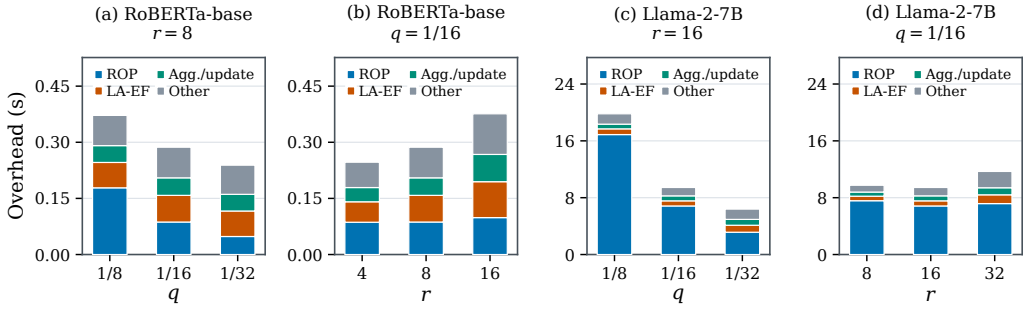


Figure 9: ProLoRA overhead breakdown under different q and r, excluding local training. Other includes serialization and remaining protocol operations.

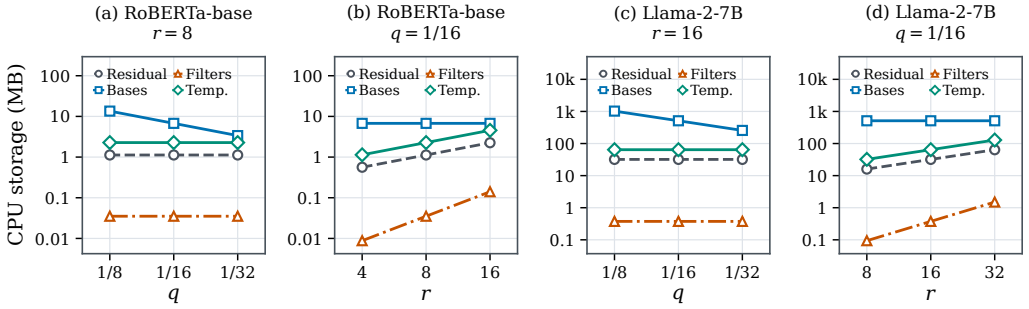


Figure 10: ProLoRA memory breakdown under different q and r (log scale). Residual storage is per client. All plotted storage is on the CPU. The native compression kernels require no GPU tensor workspace.

ProLoRA memory breakdown. Figure 10 separates residual storage, projection bases, filter caches and temporary tensor workspace. Each client stores a 1.125 MB residual for RoBERTa-base and a 32 MB residual for Llama-2-7B on the CPU. Reducing q from $1/8$ to $1/32$ decreases basis storage from 13.50 MB to 3.38 MB and from 1024 MB to 256 MB, respectively. Residual storage grows linearly with r, while the filter cache grows quadratically. Temporary tensor workspace is measured separately in the native compression kernels, excluding inputs and retained outputs. These operations run on

the CPU and allocate no GPU tensors. Their storage is separate from the training GPU memory in Table 5.

E.4 Effect of the degree of non-IID Data

In the main manuscript, experiments are primarily conducted under the Dir($\alpha = 0.5$) setting. To further explore the impact of data heterogeneity, we supplement additional comparative experiments on the QNLI and MNLI datasets under IID data and Dir($\alpha = 0.3$). A smaller α value indicates higher data heterogeneity. Other key experimental parameters remain unchanged, namely $N = 10$, $T = 300$, $r = 8$, and $q = 1/16$.

Table 6: Performance (%) comparison on QNLI and MNLI datasets under IID data and different degrees α of non-IID data.

Method	QNLI			MNLI		
	IID	$\alpha = 0.5$	$\alpha = 0.3$	IID	$\alpha = 0.5$	$\alpha = 0.3$
FedIT	91.99±0.19	**91.49**±0.15	**91.14**±0.86	85.30±0.02	**84.67**±0.18	**83.74**±0.12
FFA-LoRA	90.21±0.26	88.64±1.17	88.08±1.01	81.91±0.01	80.19±0.37	75.30±0.47
RoLoRA	91.41±0.03	90.45±0.66	89.67±1.38	84.15±0.03	83.32±0.03	78.96±4.11
Fed-SB	87.74±0.27	85.69±1.16	84.80±1.79	72.86±1.51	54.89±14.44	40.74±5.13
EcoLoRA	92.09±0.13	91.22±0.54	89.91±0.64	84.88±0.03	83.21±1.30	78.84±0.53
FedSRD	89.03±0.00	87.56±0.89	86.54±0.96	80.35±0.01	77.95±1.59	71.99±4.26
ProLoRA	**92.55**±0.14	91.40±0.83	90.62±1.10	**85.35**±0.02	84.15±0.51	79.84±2.82

Table 6 presents the comparative results on the QNLI and MNLI datasets. We observe that as data heterogeneity increases, the performance of FedIT gradually decreases, and the performance of all communication-efficient baseline methods also declines. ProLoRA achieves the highest accuracy under IID data on both datasets. At $\alpha = 0.3$, ProLoRA achieves 90.62% on QNLI, 0.52% below FedIT.

E.5 Effect of the Number of Clients

In the main manuscript, experiments are primarily conducted with $N = 10$. To further investigate the impact of the number of clients, we vary the client count to 5 and 20, and conduct experiments on the QNLI and MNLI datasets. Other key experimental parameters remain consistent with the main manuscript, namely $T = 300$, $r = 8$, $\alpha = 0.5$, and $q = 1/16$.

Table 7: Performance (%) comparison on QNLI and MNLI datasets with different numbers N of clients.

Method	QNLI			MNLI		
	$N = 5$	$N = 10$	$N = 20$	$N = 5$	$N = 10$	$N = 20$
FedIT	**91.58**±0.81	**91.49**±0.15	**91.55**±0.34	**84.50**±0.30	**84.67**±0.18	**84.76**±0.15
FFA-LoRA	89.68±1.13	88.64±1.17	87.34±1.21	81.12±0.27	80.19±0.37	78.38±0.33
RoLoRA	90.56±0.12	90.45±0.66	90.34±0.24	83.62±0.17	83.32±0.03	82.97±0.53
Fed-SB	87.45±1.07	85.69±1.16	83.28±1.24	62.03±10.71	54.89±14.44	59.24±12.13
EcoLoRA	91.26±0.18	91.22±0.54	90.89±0.23	82.68±0.61	83.21±1.30	82.13±0.57
FedSRD	88.31±0.16	87.56±0.89	86.27±0.33	78.31±0.21	77.95±1.59	79.26±1.37
ProLoRA	91.54±0.88	91.40±0.83	91.40±0.27	84.06±0.96	84.15±0.51	84.52±0.08

Table 7 presents the results with different numbers of clients. As the number of clients increases from 5 to 20, several communication-efficient baselines exhibit varying degrees of performance degradation, while ProLoRA remains stable on both QNLI and MNLI and achieves the second-best accuracy in all settings. Specifically, the performance gap between ProLoRA and FedIT remains within 0.15% on QNLI and 0.52% on MNLI, showing that ProLoRA maintains performance close to full-precision federated LoRA across different numbers of clients.

E.6 Effect of the Rank of LoRA

In the main manuscript, we primarily conduct experiments with a LoRA rank $r = 8$. We further explore the impact of different LoRA ranks, with $r = 4$ and $r = 16$. The corresponding ranks of Fed-SB are 20, 28 and 39 for $r = 4, 8, 16$, respectively, following the same communication budget matching rule. We conduct experiments on the QNLI and MNLI datasets. Other key experimental parameters remain consistent with the main manuscript, namely $N = 10$, $T = 300$, $\alpha = 0.5$, and $q = 1/16$.

Table 8: Performance (%) comparison on QNLI and MNLI datasets with different ranks r of LoRA.

Method	QNLI			MNLI		
	$r = 4$	$r = 8$	$r = 16$	$r = 4$	$r = 8$	$r = 16$
FedIT	**90.84**±0.09	**91.49**±0.15	**91.74**±0.35	**83.80**±0.24	**84.67**±0.18	**85.17**±0.19
FFA-LoRA	87.11±0.06	88.64±1.17	90.50±0.09	76.96±1.56	80.19±0.37	82.10±0.41
RoLoRA	89.38±0.47	90.45±0.66	91.06±0.45	81.93±0.53	83.32±0.03	84.15±0.05
Fed-SB	83.93±1.43	85.69±1.16	87.15±0.88	49.02±6.99	54.89±14.44	64.09±11.57
EcoLoRA	90.45±0.34	91.22±0.54	91.28±0.12	82.84±1.14	83.21±1.30	83.26±2.09
FedSRD	86.66±0.54	87.56±0.89	88.28±0.60	73.85±5.29	77.95±1.59	77.64±1.83
ProLoRA	90.43±1.16	91.40±0.83	91.73±0.03	82.23±1.44	84.15±0.51	85.15±0.07

The corresponding results are reported in Table 8. The results show that larger LoRA ranks generally lead to higher performance. However, this comes at the cost of increased communication overhead. As r increases from 4 to 16, the accuracy of ProLoRA improves from 90.43% to 91.73% on QNLI and from 82.23% to 85.15% on MNLI. ProLoRA remains within 0.52% of FedIT at $r = 8$ and $r = 16$.

F Implementation Details

F.1 Details on Hyperparameters

Unless otherwise specified, the experimental hyperparameters used in this paper are summarized in Table 9.

Table 9: The experimental hyperparameters used in this work.

Hyperparameter	RoBERTa-base	Llama-2-7B	Gemma-7B
Clients number N	10	10	10
Batch size	32	8	8
Local learning rate η_l	learning rate search from $\{1e-3, 5e-4, 1e-4, 5e-5, 2e-5\}$		
Global learning rate η_g	learning rate search from $\{1, 1.5, 2\}$		
Communication round T	300	600	600
Local update step E	20	10	10
Dirichlet parameter α	0.5	0.5	0.5
Local LoRA rank r	8	16	16
LoRA dropout rate	0.05	0.1	0.1
Target module	["query", "value"]	["q_proj", "v_proj"]	["q_proj", "v_proj"]

F.2 Details on Datasets and Benchmarks

GLUE Benchmark. GLUE (Wang et al., 2018) is a widely recognized benchmark designed to evaluate the natural language understanding capabilities of language models. We utilize five of its datasets, namely MRPC, SST-2, QNLI, QQP, MNLI.

- **MRPC** focuses on the paraphrase identification task, which requires models to judge whether two given sentences are semantically equivalent.

- **SST-2** is a sentiment classification benchmark built on movie review corpora and short text samples, where the core objective is to distinguish between positive and negative sentiment orientations in input sentences.
- **QNLI** is designed for natural language inference in the question-answering scenario, tasked with verifying whether a provided context sentence can serve as a valid answer to the corresponding question.
- **QQP** targets the task of question similarity matching, evaluating a model's capability to recognize whether two distinct questions express identical underlying intentions.
- **MNLI** is a classic natural language inference benchmark that consists of two distinct variants: matched and mismatched. It mandates models to categorize the logical relationship between a premise and a hypothesis into three distinct labels: entailment, contradiction, and neutrality.

Dolly-15k Dataset. Databricks-Dolly-15k (Zhang et al., 2024) is an open-source dataset of instruction-following records, generated by thousands of Databricks employees across diverse behavioral categories, such as brainstorming, classification, closed QA, generation, information extraction, open QA and summarization.

MetaMathQA Dataset. MetaMathQA (Yu et al., 2023) is a comprehensive and high-quality mathematical reasoning dataset consisting of 395K forward and reverse math question-answer pairs generated by large language models.

MMLU Benchmark. MMLU (Hendrycks et al., 2020) is a comprehensive benchmark designed to evaluate general-purpose language understanding across 57 diverse subjects, ranging from humanities and social sciences to STEM fields.

DROP Benchmark. DROP (Dua et al., 2019) focuses on discrete reasoning tasks, requiring models to extract numerical information and perform arithmetic operations based on given paragraphs.

GSM8K Benchmark. GSM8K (Cobbe et al., 2021) is a dataset of 8,000 grade-school level math word problems, targeting models' ability to solve multi-step arithmetic problems with natural language reasoning.

MATH Benchmark. MATH (Hendrycks et al., 2021) is a challenging benchmark consisting of advanced math problems covering algebra, calculus, geometry and other high-level topics, intended to assess the complex reasoning capabilities of language models.