SEFORA: SKETCH-AGGREGATED FEDERATED LORA WITH HETEROGENEOUS CLIENT RANKS

Anonymous authors
Paper under double-blind review

ABSTRACT

We consider federated parameter-efficient fine-tuning of foundation models, e.g., LLMs and LVMs, using low-rank adaptation (LoRA; Hu et al. 2022). Combining LoRA with federated learning introduces challenges absent from either setting alone: clients may use different LoRA ranks, making their factor matrices dimension-incompatible, and factor-wise averaging suffers from a bilinear mismatch. We propose SEFORA, a sketch-aggregated federated LoRA algorithm in which each client transmits a linear sketch of its local $\Delta\mathbf{W}$ updates, enabling direct aggregation at the federator. As a result, SEFORA overcomes the rank heterogeneity challenge and alleviates the bilinear mismatch. We establish convergence to a neighborhood of a first-order stationary point at rate $\mathcal{O}(T^{-1/2})$. Extensive numerical experiments on language and vision models showcase the performance of our algorithm compared to existing methods.

1 INTRODUCTION

Foundation models have remarkable learning capabilities and demonstrate strong performance across a diverse range of tasks. Adapting such foundation models to downstream tasks relies on access to vast, potentially heterogeneous datasets and requires tremendous computational resources, especially when fine-tuning all model parameters. Since fine-tuning all model parameters requires substantial computational and memory resources, parameter-efficient fine-tuning (PEFT) methods, which adapt a pretrained model by updating only a small, structured subset of parameters, are employed, see (Houlsby et al., 2019; Liao et al., 2023; Hu et al., 2022; Tian et al., 2025; Han et al., 2024).

Training on large amounts of distributed data can be achieved using federated learning (FL) (McMahan et al., 2017). FL is a distributed learning paradigm in which a central server (federator) coordinates the training of a neural network on heterogeneous data generated and owned by participating clients. The clients keep their data, train the neural network locally, and share only their model updates with the federator. As such, FL enhances data privacy by keeping raw data local, reduces communication and storage costs associated with transferring data to a central server, and scales well due to its ease of implementation at the clients. FL has attracted significant attention from the scientific community and is being applied across several domains (Kairouz & McMahan, 2021; Jiang et al., 2025).

Combining PEFT with FL spans adapter-based methods (Cai et al., 2023; Ghiasvand et al., 2024), prompt tuning (Zhao et al., 2023; Qiu et al., 2023), selective parameter optimization (Yu et al., 2023), and low-rank gradient subspace optimization (Peng et al., 2026). We focus on federated PEFT methods using Low-Rank Adaptation (LoRA) (Hu et al., 2022), which has emerged as one of the most popular PEFT techniques, e.g., (Tian et al., 2025; Zhang et al., 2023; Hayou et al., 2024; Liu et al., 2024c; Zhou et al., 2025a; Chen et al., 2024b; Kim et al., 2025) and used in FL, e.g., (Zhang et al., 2024; Liu et al., 2026b;a; Cho et al., 2024; Chen et al., 2024a; Sun et al., 2024).

LoRA-based FL introduces challenges that do not arise in either centralized LoRA or conventional FL alone. Unlike standard FL, LoRA updates cannot be averaged directly. In LoRA, for each layer of the network, each client i computes an update of the form[1] $\Delta\mathbf{W}_i = \mathbf{B}_i\mathbf{A}_i$, where the model update $\Delta\mathbf{W}_i \in \mathbb{R}^{d\times s}$ is represented as the product of two factor matrices called *adapters* $\mathbf{B}_i \in \mathbb{R}^{d\times r}$, $\mathbf{A}_i \in \mathbb{R}^{r\times s}$, for a chosen $r \ll \min(d, s)$. This factorization, together with the choice of r, reduces the

[1]We omit the layer index as the operation is the same for all layers.

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.

number of trainable parameters per layer from ds to $(d+s)r$ and restricts the rank of $\Delta\mathbf{W}_i$ to r. In LoRA-based FL, the federator aims to obtain $\Delta\mathbf{W} \triangleq \sum_i p_i \Delta\mathbf{W}_i = \sum_i p_i \mathbf{B}_i\mathbf{A}_i$, where $p_i \geq 0$ and $\sum_i p_i = 1$. Transmitting $\Delta\mathbf{W}_i$ to the federator is undesirable, as it incurs the same communication cost as transmitting a full fine-tuning update. Instead, the clients can transmit $\mathbf{B}_i$ and $\mathbf{A}_i$ separately. The federator has two options: *(i)* compute each $\Delta\mathbf{W}_i = \mathbf{B}_i\mathbf{A}_i$ and average them to obtain $\Delta\mathbf{W}$; or *(ii)* average the $\mathbf{A}_i$'s and $\mathbf{B}_i$'s first to obtain $\overline{\mathbf{A}} = \sum_i p_i\mathbf{A}_i$ and $\overline{\mathbf{B}} = \sum_i p_i\mathbf{B}_i$ and compute $\overline{\Delta\mathbf{W}} = \overline{\mathbf{B}}\,\overline{\mathbf{A}}$. Each approach introduces its own challenges, affecting either the computational efficiency or the accuracy of the aggregated update, which in turn influences the convergence of the federated fine-tuning algorithm. We focus on the following challenges.

Challenge 1: Rank-heterogeneity In the rank-heterogeneous setting, clients use different adapter ranks r_i to match their local compute budgets or the complexity of their data (Hu et al., 2022), resulting in matrices with incompatible dimensions in aggregation.

Challenge 2: Bilinear mismatch Zero-padding and averaging the adapters first leads to the *bilinear mismatch* problem (also known as aggregation bias): $\overline{\mathbf{B}}\,\overline{\mathbf{A}} \neq \sum_i p_i \mathbf{B}_i\mathbf{A}_i$. The bilinear mismatch problem is exacerbated when considering differential privacy (Liu et al., 2026a) and the non-linearity does not allow for secure aggregation, used in privacy-preserving FL (Kairouz & McMahan, 2021).

Closely related work We summarize the closest work and give an extensive review in Appendix D.

Rank heterogeneity is addressed either by operating in the full model space or by padding the adapters to the same dimensions. In FLoRA (Wang et al., 2024), the federator stacks the adapters and sends them to the clients; thus increasing the federator-to-client communication cost. A key difference is that clients update the backbone model and reinitialize their adapters at each iteration. In FlexLoRA (Bai et al., 2024), the federator computes and averages the updates $\Delta\mathbf{W}_i = \mathbf{B}_i\mathbf{A}_i$ to obtain $\Delta\mathbf{W}$ and applies SVD on $\Delta\mathbf{W}$ with truncation to obtain rank-r_i update factors sent to client i; incurring an $\mathcal{O}(Ndrs)$ computational overhead per layer for computing the $\Delta\mathbf{W}_i$'s and $\mathcal{O}(d^2s)$ for the SVD. In FedHL (Peng et al., 2025), the federator also computes the dense updates $\Delta\mathbf{W}_i = \mathbf{B}_i\mathbf{A}_i$ and applies SVD truncation like FlexLoRA, but it introduces a compensation aggregation rule and dynamically adjusted aggregation weights to mitigate the biases caused by truncating the global model to different local ranks. In (Cho et al., 2024), the authors remark that computing $\Delta\mathbf{W}_i$ first loses information on the cross-relation across clients and corroborate this observation through numerical experiments. Hence, in their HetLoRA, the federator zero-pads the clients' $\mathbf{A}_i$'s and $\mathbf{B}_i$'s and aggregates them using a careful weighted average. The computational complexity of computing the weights and the global update is $\mathcal{O}(Ndrs)$. Nevertheless, this method does not mitigate the bilinear mismatch problem. Furthermore, zero-padding the matrices is not always faithful to the clients' rank representations, which is why careful computation of the p_i's is needed and why other works, e.g., (Byun & Lee, 2025; Ha et al., 2026), replace zero-padding with column repetition. In FSLoRA (Fang et al., 2026), the LoRA parametrization is modified to $\Delta\mathbf{W} = \mathbf{BSA}$, where $\mathbf{S}$ is a diagonal $r \times r$ matrix for a given global rank r. Each client draws a random diagonal $r \times r$ matrix $\mathbf{S}_i$ with only r_i non-zero entries and updates $\mathbf{B}_i\mathbf{S}_i$ and $\mathbf{S}_i\mathbf{A}_i$ with rank r_i. Notice that $\mathbf{B}_i\mathbf{S}_i$ and $\mathbf{S}_i\mathbf{A}_i$ have the same dimension across clients, enabling aggregation. The drawbacks of this method are that it does not mitigate the bilinear mismatch and requires clients to randomly choose the subspace to be updated, as dictated by the choice of $\mathbf{S}_i$, thereby incurring additional variance in convergence.

To our knowledge, there is no rank-heterogeneous LoRA-based FL that mitigates the bilinear mismatch and does not operate in the full model space. Also, only Fang et al. (2026) provide theoretical insights into the convergence of the rank-heterogeneous algorithm for non-convex loss functions.

Our contributions We introduce SeFoRA, a rank-agnostic federated LoRA-based PEFT method. It alleviates the bilinear mismatch and does not operate in the full model space. Instead of transmitting $\mathbf{A}_i$ and $\mathbf{B}_i$, each client sketches its $\Delta\mathbf{W}_i$ into two fixed-size matrices using the linear sketching method in (Tropp et al., 2017). The size of the sketch is comparable to the size of $\mathbf{B}_i$ and $\mathbf{A}_i$. Since the sketching operator is linear, aggregating the sketches at the federator is equivalent to computing the sketch of the aggregated update $\Delta\mathbf{W}$. The main properties of SeFoRA are summarized below.

1. SeFoRA aggregates fixed-dimensional sketches, irrespective of clients' local ranks, and can be seamlessly combined with adaptive rank-truncation at the clients. It converges at rate

$\mathcal{O}(T^{-1/2})$ to a neighborhood of a first-order stationary point characterized by sketching error, rank truncation, and client heterogeneity.

2. Numerical experiments on fine-tuning language and vision models in the rank-heterogeneous setting showcase the performance of SeFoRA compared to existing methods.
3. Mitigating the bilinear mismatch is particularly desirable in settings where differential privacy is required, and linear aggregation enables secure aggregation. Since this work is not directly concerned with privacy guarantees, we will explore those benefits in future work.

2 Setup and Problem Formulation

For an integer N, let $[N] = \{1, \cdots, N\}$. Vectors and matrices are denoted by bold lowercase and uppercase letters, respectively. For a matrix $\mathbf{A}$ with rank r, we denote its singular values as $\sigma_1(\mathbf{A}) \geq \cdots \geq \sigma_r(\mathbf{A}) \geq 0$. Let $\|\mathbf{A}\|_2 = \sigma_1(\mathbf{A})$ denote the spectral norm and $\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{r} \sigma_i(\mathbf{A})^2}$ the Frobenius norm. For a matrix $\mathbf{A}$, $\mathbf{A}_{:,1:r}$ and $\mathbf{A}_{1:r,:}$ denote the submatrices formed by its first r columns and first r rows, respectively, and $\mathbf{A}_{1:r,1:r}$ denotes its leading $r \times r$ submatrix. For a tuple of matrices $\mathtt{B} = (\mathbf{B}_1, \cdots, \mathbf{B}_L)$, we let $\|\mathtt{B}\|_2 \triangleq \max_{l\in[L]} \|\mathbf{B}_l\|_2$, $\|\mathtt{B}\|_F \triangleq \sqrt{\sum_l \|\mathbf{B}_l\|_F^2}$, and we say that $\mathtt{B}$ has rank r if $\mathrm{rank}(\mathbf{B}_l) \leq r$ for every $l \in [L]$. The transpose of a tuple $\mathtt{B}$ is denoted by $\mathtt{B}^\top \triangleq ((\mathbf{B}_1)^\top, \cdots, (\mathbf{B}_L)^\top)$. The product and addition of two tuples of matrices $\mathtt{A}$ and $\mathtt{B}$ are defined as $\mathtt{BA} = (\mathbf{B}_1\mathbf{A}_1, \cdots, \mathbf{B}_L\mathbf{A}_L)$ and $\mathtt{B} + \mathtt{A} = (\mathbf{B}_1 + \mathbf{A}_1, \cdots, \mathbf{B}_L + \mathbf{A}_L)$, respectively. The inner product of two tuples $\mathtt{A}$ and $\mathtt{B}$ is $\langle \mathtt{A}, \mathtt{B}\rangle \triangleq \sum_{l=1}^{L} \langle \mathbf{A}_l, \mathbf{B}_l\rangle = \sum_{l=1}^{L} \mathrm{tr}((\mathbf{A}_l)^\top \mathbf{B}_l)$. Let $\mathrm{QR}(\cdot)$ denote the reduced QR factorization, and $\mathrm{SVD}(\cdot)$ denote the compact SVD decomposition.

Learning Objective We consider an FL system consisting of a federator and N clients, indexed by $i \in [N]$. Client i holds a local dataset $\mathcal{D}_i$, with $|\mathcal{D}_i| > 0$ data samples. Define $\mathcal{D} = \bigcup_{i\in[N]} \mathcal{D}_i$. Let a tuple of matrices $\mathtt{W} = (\mathbf{W}_1, \cdots, \mathbf{W}_L)$ denote the L linear layers selected for fine-tuning, where for each layer $\mathbf{W}_l \in \mathbb{R}^{d_l \times s_l}$. Let $\mathtt{W}_0$ denote the pre-trained values of $\mathtt{W}$ and $\Delta\mathtt{W}$ be the change of $\mathtt{W}$ during fine-tuning, such that $\mathtt{W} = \mathtt{W}_0 + \Delta\mathtt{W}$. For a sample $(\boldsymbol{x}, y)$, let $f(\mathtt{W}; \boldsymbol{x})$ denote the output of a neural network parameterized by $\mathtt{W}$, and $\ell(f(\mathtt{W}; \boldsymbol{x}), y)$ be the sample-wise loss function. The local empirical loss of client i is defined as

$$\mathcal{L}_i^{\text{full}}(\Delta\mathtt{W}_i; \mathcal{D}_i) \triangleq \frac{1}{|\mathcal{D}_i|} \sum_{(\boldsymbol{x},y)\in\mathcal{D}_i} \ell(f(\mathtt{W}_0 + \Delta\mathtt{W}_i; \boldsymbol{x}), y).$$

Here, "full" refers to an unfactorized update of the selected fine-tuning layers. Let $p_i > 0$ with $\sum_{i=1}^{N} p_i = 1$ be the aggregation weight assigned to client i. The global learning objective is

$$\mathcal{L}^{\text{full}}(\Delta\mathtt{W}; \mathcal{D}) \triangleq \sum_{i=1}^{N} p_i \mathcal{L}_i^{\text{full}}(\Delta\mathtt{W}; \mathcal{D}_i).$$

Fine-tuning and LoRA In LoRA, $\mathtt{W}_0$ is frozen, i.e., it does not receive gradient updates during fine-tuning. For each layer $l \in [L]$, LoRA represents the update $\Delta\mathbf{W}_l$ as the product of two factor matrices of smaller dimensions $\Delta\mathbf{W}_l = \mathbf{B}_l\mathbf{A}_l$, where $\mathbf{B}_l \in \mathbb{R}^{d_l \times r_l}$, and $\mathbf{A}_l \in \mathbb{R}^{r_l \times s_l}$, and $r_l \ll \min(d_l, s_l)$. We denote the tuple of low-rank factor matrices for all the layers as $\mathtt{B} = (\mathbf{B}_1, \cdots, \mathbf{B}_L)$ and $\mathtt{A} = (\mathbf{A}_1, \cdots, \mathbf{A}_L)$, thus, $\Delta\mathtt{W} = \mathtt{BA}$.

The local LoRA objective is the full update objective evaluated under this factorization:

$$\mathcal{L}_i^{\text{lora}}(\mathtt{B}_i, \mathtt{A}_i; \mathcal{D}_i) \triangleq \mathcal{L}_i^{\text{full}}(\mathtt{B}_i\mathtt{A}_i; \mathcal{D}_i).$$

Similarly, the global LoRA objective is

$$\mathcal{L}^{\text{lora}}(\mathtt{B}, \mathtt{A}; \mathcal{D}) \triangleq \sum_{i=1}^{N} p_i \mathcal{L}_i^{\text{lora}}(\mathtt{B}, \mathtt{A}; \mathcal{D}_i) = \sum_{i=1}^{N} p_i \mathcal{L}_i^{\text{full}}(\mathtt{BA}; \mathcal{D}_i) = \mathcal{L}^{\text{full}}(\mathtt{BA}; \mathcal{D}).$$

Federated LoRA Client i uses the local rank profile $\mathbf{r}_i = (r_{i,1}, \cdots, r_{i,L})$ with local factors $\mathtt{B}_i = (\mathbf{B}_{i,1}, \cdots, \mathbf{B}_{i,L})$ and $\mathtt{A}_i = (\mathbf{A}_{i,1}, \cdots, \mathbf{A}_{i,L})$, where $\mathbf{B}_{i,l} \in \mathbb{R}^{d_l \times r_{i,l}}$ and $\mathbf{A}_{i,l} \in \mathbb{R}^{r_{i,l} \times s_l}$. The

ranks may differ across clients and layers according to local computational and memory constraints. Define the global rank profile $\mathbf{r} = (r_1, \cdots, r_L)$ as $r_l = \max_{i \in [N]} r_{i,l}$ for all $l \in [L]$.

The training proceeds for T communication rounds. At round $t \in \{0, \cdots, T-1\}$, the federator holds the current global adapters $(\mathtt{B}^{(t)}, \mathtt{A}^{(t)})$, constructs initial client-specific local adapters $(\mathtt{B}_i^{(t)}, \mathtt{A}_i^{(t)})$ according to the client's local rank profile, and transmits these factors to client i. Client i then performs E local steps of mini-batch stochastic gradient descent (SGD) on its local LoRA objective using step size $\eta^{(t)} > 0$. We denote by $(\mathtt{B}_i^{(t,e)}, \mathtt{A}_i^{(t,e)})$ the local parameters at step $e \in \{0, \cdots, E-1\}$ during communication round t. Specifically, at each local step e, client i samples a mini-batch of data $\xi_i^{(t,e)} \subset \mathcal{D}_i$ and computes the stochastic gradient of the local objective. The local update rules are given by $\mathtt{B}_i^{(t,e+1)} = \mathtt{B}_i^{(t,e)} - \eta^{(t)} \nabla_{\mathtt{B}} \mathcal{L}_i^{\text{lora}}(\mathtt{B}_i^{(t,e)}, \mathtt{A}_i^{(t,e)}; \xi_i^{(t,e)})$, $\mathtt{A}_i^{(t,e+1)} = \mathtt{A}_i^{(t,e)} - \eta^{(t)} \nabla_{\mathtt{A}} \mathcal{L}_i^{\text{lora}}(\mathtt{B}_i^{(t,e)}, \mathtt{A}_i^{(t,e)}; \xi_i^{(t,e)})$, where $\mathtt{B}_i^{(t,0)} = \mathtt{B}_i^{(t)}$ and $\mathtt{A}_i^{(t,0)} = \mathtt{A}_i^{(t)}$. After E local steps, clients upload their updates to the federator.

Challenges of LoRA with FedAvg Upon receiving the clients' local updates, the federator needs to compute a global $\mathtt{B}^{(t+1)}$ and $\mathtt{A}^{(t+1)}$. In rank-homogeneous settings where all clients have the same rank for each layer, the federator could use vanilla FedAvg as in FedIT (Zhang et al., 2024). However, in rank-heterogeneous settings, the adapters have different dimensions and averaging is not directly possible. Existing approaches either pad the adapters to make them addition-compatible or operate in the full model space. The former faces the bilinear mismatch problem and the faithful representation of the clients' ranks, whereas the latter incurs high computational cost at the federator. The question that remains unanswered in the literature and is answered affirmatively in this work is:

Research question: *Is it possible to construct a rank-heterogeneous LoRA-based FL method that mitigates the bilinear mismatch and does not operate in the full model space?*

3 Our Algorithm and Contributions

We jointly tackle rank heterogeneity and the bilinear mismatch without operating in the full model space. The main ingredient of our algorithm is operating in a sketched-space of the full model using the linear matrix sketching method introduced by Tropp et al. (2017). For ease of presentation, we describe the method for a single layer. The same procedure is applied independently to each layer, and the resulting matrices are collected into tuples. Therefore, we omit the layer index l. Each client sketches (compresses) its local update $\Delta\mathbf{W}_i \in \mathbb{R}^{d \times s}$ into two matrices, $\mathbf{Y}_i \in \mathbb{R}^{d \times r}$ and $\mathbf{Z}_i \in \mathbb{R}^{k \times s}$, where $r = r_l = \max_{i \in [N]} r_{i,l}$ is the maximum local rank for each layer and k is a constant satisfying $k > r + 1$. By aggregating the $\mathbf{Y}_i$ and $\mathbf{Z}_i$ sent by the clients, the federator obtains a sketch of the global update $\Delta\mathbf{W} = \sum_{i \in [N]} p_i \Delta\mathbf{W}_i$, from which it computes a low-rank approximation of $\Delta\mathbf{W}$. Therefore, this maps local adapters of rank r_i to sketches of dimension r and k, respectively, thereby enabling linear aggregation and bypassing the bilinear mismatch while avoiding the aggregation in the full ds-dimensional space $\Delta\mathbf{W}$. This incurs an additional communication of $(r - r_i)d$ and $(k - r_i)s$, which remains low due to the choice of k. Before presenting our theoretical guarantees and comparing SEFORA numerically to the state of the art, we first explain the sketching method introduced by Tropp et al. (2017).

Sketching as Low-Rank Approximation (Tropp et al., 2017) Suppose $\Delta\mathbf{W} \in \mathbb{R}^{d \times s}$ is an arbitrary matrix. Let $r \ll \min\{d, s\}$ be the target rank. Given a parameter $k > r + 1$, let $\mathbf{\Omega} \in \mathbb{R}^{s \times r}$ and $\mathbf{\Psi} \in \mathbb{R}^{k \times d}$ be independent random matrices with entries drawn independently from the standard normal distribution. Two *sketch matrices* of $\Delta\mathbf{W}$ are produced via left and right matrix multiplication:

$$\mathbf{Y} \triangleq \Delta\mathbf{W}\mathbf{\Omega} \quad \text{and} \quad \mathbf{Z} \triangleq \mathbf{\Psi}\Delta\mathbf{W}.$$

Given $\mathbf{Y} \in \mathbb{R}^{d \times r}$, and $\mathbf{Z} \in \mathbb{R}^{k \times s}$, the low-rank approximation $\widehat{\Delta\mathbf{W}}$ of $\Delta\mathbf{W}$ with rank r is computed using the following steps: *(i)* factorize $\mathbf{Y}$ via QR decomposition: $(\widehat{\mathbf{B}}, \sim) \leftarrow \mathrm{QR}(\mathbf{Y})$; *(ii)* compute $\mathbf{\Psi}\widehat{\mathbf{B}}$ and find its QR decomposition: $(\mathbf{Q}, \mathbf{R}) \leftarrow \mathrm{QR}(\mathbf{\Psi}\widehat{\mathbf{B}})$; *(iii)* compute $\widehat{\mathbf{A}} = \mathbf{R}^{-1}(\mathbf{Q}^T\mathbf{Z})$; and, *(iv)* produce $\widehat{\mathbf{B}}$ and $\widehat{\mathbf{A}}$, which form the low-rank approximation $\widehat{\Delta\mathbf{W}} = \widehat{\mathbf{B}}\widehat{\mathbf{A}}$ of $\Delta\mathbf{W}$. This sketch-inversion procedure is summarized in Algorithm 2.

3.1 SEFORA: SKETCH-AGGREGATED FEDERATED LORA

For notational simplicity, we present the algorithm under the assumption that all adapted layers share the same input and output dimensions, i.e., $d_l = d$ and $s_l = s$. The method extends directly to heterogeneous dimensions by using compatible sizes.

Initialization The federator has the pretrained model $\mathtt{W}_0$. Each client sets a local rank profile[2] $\mathbf{r}_i = (r_{i,1}, \cdots, r_{i,L})$ based on local compute and memory constraints, i.e., $r_{i,l}$ is the maximum rank that client i can allocate to layer l. This profile is communicated to the federator. The federator creates a global rank profile $\mathbf{r} = (r_1, \cdots, r_L)$ where $r_l = \max_{i\in[N]} r_{i,l}$ for all $l \in [L]$, and chooses a sketching dimension $k_l > r_l + 1$. We further define $r_{\max} \triangleq \max_{l\in[L]} r_l$, and $k_{\max} \triangleq \max_{l\in[L]} k_l$. The rank profiles are fixed throughout training. Following standard practice (Hu et al., 2022), the federator initializes the global adapters $\mathbf{B}_l^{(0)} \in \mathbb{R}^{d\times r_l}$ as the zero matrix and $\mathbf{A}_l^{(0)} \in \mathbb{R}^{r_l\times s}$ with independent random Gaussian entries.

In the remainder, without loss of generality, we focus on a single layer l and a single round t.

Federator-side computation In round t, the federator draws two independent random matrices $\mathbf{\Omega}^{(t)} \in \mathbb{R}^{s\times r_{\max}}$ and $\mathbf{\Psi}^{(t)} \in \mathbb{R}^{k_{\max}\times d}$ from the standard normal distribution, and communicates them to the clients via a shared random seed. It does the following steps.

Client-specific truncation. To reduce the downlink communication, the federator truncates the global adapters according to the client's local rank for each client i:

$$\mathbf{B}_{i,l}^{(t)} = (\mathbf{B}_l^{(t)})_{:,1:r_{i,l}} \quad \text{and} \quad \mathbf{A}_{i,l}^{(t)} = (\mathbf{A}_l^{(t)})_{1:r_{i,l},:},$$

and broadcasts $\mathtt{B}_i^{(t)}$ and $\mathtt{A}_i^{(t)}$ to the client i.

Global-adapter sketching for usage in next round. Federator computes the layer-specific matrices $\mathbf{\Omega}_l^{(t)} = \mathbf{\Omega}_{:,1:r_l}^{(t)} \in \mathbb{R}^{s\times r_l}$ and $\mathbf{\Psi}_l^{(t)} = \mathbf{\Psi}_{1:k_l,:}^{(t)} \in \mathbb{R}^{k_l\times d}$ for all $l \in [L]$, and stores the sketches of the current global update as

$$\mathbf{Y}_{\text{glo},l}^{(t)} = \mathbf{B}_l^{(t)}(\mathbf{A}_l^{(t)}\mathbf{\Omega}_l^{(t)}) \quad \text{and} \quad \mathbf{Z}_{\text{glo},l}^{(t)} = (\mathbf{\Psi}_l^{(t)}\mathbf{B}_l^{(t)})\mathbf{A}_l^{(t)}.$$

Client per-round computation Client i receives a shared seed from which it regenerates $\mathbf{\Omega}^{(t)} \in \mathbb{R}^{s\times r_{\max}}$ and $\mathbf{\Psi}^{(t)} \in \mathbb{R}^{k_{\max}\times d}$, and computes the layer-specific matrices $\mathbf{\Omega}_l^{(t)}$ and $\mathbf{\Psi}_l^{(t)}$.

Local fine-tuning. Client i receives the adapters $\mathtt{B}_i^{(t)}$ and $\mathtt{A}_i^{(t)}$ with the correct local dimensions, initiate the local factors as $\mathbf{B}_{i,l}^{(t,0)} = \mathbf{B}_{i,l}^{(t)}$ and $\mathbf{A}_{i,l}^{(t,0)} = \mathbf{A}_{i,l}^{(t)}$, and fine-tunes them for E local iterations to produce $\mathbf{A}_{i,l}^{(t,E)}$ and $\mathbf{B}_{i,l}^{(t,E)}$.

Adapter-change sketching. Rather than sketching the final local update, client i sketches its change relative to the client-specific initialization, $\Delta\mathbf{W}_{i,l}^{(t,E)} - \Delta\mathbf{W}_{i,l}^{(t,0)} = \mathbf{B}_{i,l}^{(t,E)}\mathbf{A}_{i,l}^{(t,E)} - \mathbf{B}_{i,l}^{(t,0)}\mathbf{A}_{i,l}^{(t,0)}$, i.e.,

$$\mathbf{Y}_{i,l}^{(t)} = \mathbf{B}_{i,l}^{(t,E)}(\mathbf{A}_{i,l}^{(t,E)}\mathbf{\Omega}_l^{(t)}) - \mathbf{B}_{i,l}^{(t,0)}(\mathbf{A}_{i,l}^{(t,0)}\mathbf{\Omega}_l^{(t)}) \in \mathbb{R}^{d\times r_l},$$
$$\mathbf{Z}_{i,l}^{(t)} = (\mathbf{\Psi}_l^{(t)}\mathbf{B}_{i,l}^{(t,E)})\mathbf{A}_{i,l}^{(t,E)} - (\mathbf{\Psi}_l^{(t)}\mathbf{B}_{i,l}^{(t,0)})\mathbf{A}_{i,l}^{(t,0)} \in \mathbb{R}^{k_l\times s},$$

by first computing the multiplication in parentheses to reduce the computation cost. The client sends $\mathbf{Y}_{i,l}^{(t)}$ and $\mathbf{Z}_{i,l}^{(t)}$ to the federator incurring a communication cost of $dr_l + k_l s$ per layer, which is low since k_l can be chosen to be $2r_l + 1$ (Tropp et al., 2017). For clients with $r_{i,l} \ll r_l$, the clients can opt to communicate the adapters to the federator, who computes the sketches.

[2] SEFORA is agnostic to how the clients adjust their local ranks. They may use adaptive techniques such as those proposed in (Yan et al., 2026a; Wu et al., 2026a), adapting the ranks based on their computational and memory resources as well as the contribution of each layer to fine-tuning in terms of effective rank. We note that adaptive LoRA rank allocation has been extensively studied in the centralized LoRA setting, e.g., (Zhang et al., 2023; Valipour et al., 2023; Ding et al., 2023; Liu et al., 2026b), and can be used in the federated LoRA setting.

Federator aggregation and communication Due to the linearity of the sketching operation, aggregating $\mathbf{Y}_{i,l}^{(t)}$ and $\mathbf{Z}_{i,l}^{(t)}$ is implicitly aggregating the client local changes $\Delta\mathbf{W}_{i,l}^{(t,E)} - \Delta\mathbf{W}_{i,l}^{(t,0)}$, but in a smaller subspace. Hence, the federator performs the steps below.

Aggregation. The federator adds the aggregated client-change sketches to the global sketches:

$$\mathbf{Y}_l^{(t+1)} = \mathbf{Y}_{\text{glo},l}^{(t)} + \sum_{i\in[N]} p_i \mathbf{Y}_{i,l}^{(t)} \quad \text{and} \quad \mathbf{Z}_l^{(t+1)} = \mathbf{Z}_{\text{glo},l}^{(t)} + \sum_{i\in[N]} p_i \mathbf{Z}_{i,l}^{(t)}.$$

Since clients start from different rank-truncated versions of the global adapter, this updating approach keeps all clients at the same federated-averaging reference point.

Unsketching. The federator uses $\mathbf{Y}_l^{(t+1)}$ and $\mathbf{Z}_l^{(t+1)}$ as inputs for Algorithm 2 to obtain $\widehat{\mathbf{B}}_l^{(t+1)}$ and $\widehat{\mathbf{A}}_l^{(t+1)}$, i.e., $(\widehat{\mathbf{B}}_l^{(t+1)}, \widehat{\mathbf{A}}_l^{(t+1)}) \leftarrow \textsc{Unsketching}(\mathbf{Y}_l^{(t+1)}, \mathbf{Z}_l^{(t+1)}, \mathbf{\Psi}_l^{(t)})$. Then, it computes $(\mathbf{U}_l^{(t+1)}, \mathbf{\Sigma}_l^{(t+1)}, \mathbf{V}_l^{(t+1)}) \leftarrow \text{SVD}(\widehat{\mathbf{A}}_l^{(t+1)})$ and updates the global adapters as

$$\mathbf{B}_l^{(t+1)} = \widehat{\mathbf{B}}_l^{(t+1)} \mathbf{U}_l^{(t+1)} (\mathbf{\Sigma}_l^{(t+1)})^{\frac{1}{2}} \quad \text{and} \quad \mathbf{A}_l^{(t+1)} = (\mathbf{\Sigma}_l^{(t+1)})^{\frac{1}{2}} (\mathbf{V}_l^{(t+1)})^\top.$$

To initiate the next round, the federator starts with a fresh shared seed for the new random matrices, and sketches and truncates the updated global adapter. The procedure is summarized in Algorithm 1.

4 Theoretical Analysis and Insights

This section analyzes the proposed method theoretically. We use the tuple notations introduced in Section 2 throughout. We first state the smoothness and stochastic gradient assumptions used below.

Assumption 1 (Per-sample Lipschitz smoothness)**.** *There exists a $\mu > 0$ such that for any $\Delta\mathtt{W}$, $\Delta\mathtt{W}'$:*

$$\|\nabla_{\Delta\mathtt{W}} \ell(f(\mathtt{W}_0 + \Delta\mathtt{W}; \boldsymbol{x}), y) - \nabla_{\Delta\mathtt{W}} \ell(f(\mathtt{W}_0 + \Delta\mathtt{W}'; \boldsymbol{x}), y)\|_F \le \mu \|\Delta\mathtt{W} - \Delta\mathtt{W}'\|_F.$$

Assumption 2 (Unbiased stochastic gradients with bounded moments)**.** *For client i and all $\Delta\mathtt{W} \in \prod_{l=1}^{L} \mathbb{R}^{d\times s}$, the stochastic gradient $\nabla_{\Delta\mathtt{W}} \mathcal{L}_i^{\textit{full}}(\Delta\mathtt{W}; \xi_i)$ is an unbiased estimator of $\nabla_{\Delta\mathtt{W}} \mathcal{L}_i^{\textit{full}}(\Delta\mathtt{W})$, i.e.,*

$$\mathbb{E}_{\xi_i}\left[\nabla_{\Delta\mathtt{W}} \mathcal{L}_i^{\textit{full}}(\Delta\mathtt{W}; \xi_i)\right] = \nabla_{\Delta\mathtt{W}} \mathcal{L}_i^{\textit{full}}(\Delta\mathtt{W}).$$

Moreover, there are constants $\chi, \kappa > 0$ such that

$$\mathbb{E}_{\xi_i}\left[\|\nabla_{\Delta\mathtt{W}} \mathcal{L}_i^{\textit{full}}(\Delta\mathtt{W}; \xi_i)\|_F^2\right] \le \chi^2, \quad \textit{and} \quad \mathbb{E}_{\xi_i}\left[\|\nabla_{\Delta\mathtt{W}} \mathcal{L}_i^{\textit{full}}(\Delta\mathtt{W}; \xi_i)\|_F^4\right] \le \kappa^4.$$

Assumption 3 (Bounded Client Heterogeneity)**.** *There exists a real value ζ such that for all $\Delta\mathtt{W} = \mathtt{BA}$,*

$$\sum_{i\in[N]} p_i \|\nabla_{\Delta\mathtt{W}} \mathcal{L}_i^{\textit{full}}(\Delta\mathtt{W}) - \nabla_{\Delta\mathtt{W}} \mathcal{L}^{\textit{full}}(\Delta\mathtt{W})\|_F^2 \le \zeta^2.$$

To facilitate the convergence analysis, we adopt the following bounded-factor assumption, which is also used in prior theoretical analyses of federated LoRA (Guo et al., 2025; Chen et al., 2026).

Assumption 4 (Bounded LoRA factors)**.** *There exist constants $M_{\mathbf{A}}, M_{\mathbf{B}} > 0$ such that, for every layer l, round t, client i, and local step e, the spectral norms of the local and global LoRA factors are bounded, i.e., $\|\mathbf{A}_{i,l}^{(t,e)}\|_2 \le M_{\mathbf{A}}$, $\|\mathbf{B}_{i,l}^{(t,e)}\|_2 \le M_{\mathbf{B}}$, $\|\mathbf{A}_l^{(t)}\|_2 \le M_{\mathbf{A}}$, and $\|\mathbf{B}_l^{(t)}\|_2 \le M_{\mathbf{B}}$.*

Lemma 1 quantifies the error incurred by reconstructing the aggregated update from its sketches.

Lemma 1 (Low-Rank Approximation (Sketching) Error (Tropp et al., 2017))**.** *Let $\mathbf{W} \in \mathbb{R}^{d\times s}$, $r \ge 2$, $\varrho \in \{0, \cdots, r-2\}$, and $k > r+1$. Draw random matrices $\mathbf{\Omega} \in \mathbb{R}^{s\times r}$ and $\mathbf{\Psi} \in \mathbb{R}^{k\times d}$ independently from the standard normal distribution. The rank-r approximation $\widehat{\mathbf{W}}$ satisfies*

$$\mathbb{E}\|\widehat{\mathbf{W}} - \mathbf{W}\|_F^2 \le (1 + \frac{r}{k-r-1}) \cdot \min_{\varrho<r-1} (1 + \frac{\varrho}{r-\varrho-1}) \cdot \tau_{\varrho+1}^2(\mathbf{W}),$$

where $\tau_j^2(\mathbf{W}) \triangleq \min_{\textit{rank}(\mathbf{W}')<j} \|\mathbf{W} - \mathbf{W}'\|_F^2 = \sum_{i\ge j} \sigma_i^2(\mathbf{W})$ is the j-th tail energy of $\mathbf{W}$.

For a tuple of layerwise updates, Lemma 1 is applied independently to each layer.

Let $\mathcal{F}_t$ be the algorithmic history at the beginning of round t, and write $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_t]$. Define constants $C_1 \triangleq M_{\mathbf{B}}^2 + M_{\mathbf{A}}^2$, $C_2 \triangleq \chi M_{\mathbf{A}} M_{\mathbf{B}} \kappa^2 + 2\mu(M_{\mathbf{A}}^4 + M_{\mathbf{B}}^4)\chi^2$, $C_3 \triangleq \chi^2\Big(\chi^2(M_{\mathbf{A}}^2 + M_{\mathbf{B}}^2) + 2\mu^2(M_{\mathbf{A}}^2 + M_{\mathbf{B}}^2)(M_{\mathbf{B}}^4 + M_{\mathbf{A}}^4)\Big)$, and $C_4 \triangleq \mu M_{\mathbf{A}}^2 M_{\mathbf{B}}^2 \kappa^4$. Define the sketching error as $\mathcal{E}_{\text{sketch}}^{(t+1)} \triangleq \Delta \mathtt{W}_{\mathtt{SeFoRA}}^{(t+1)} - \mathtt{V}^{(t+1)}$, where $\mathtt{V}^{(t+1)} \triangleq \Delta \mathtt{W}^{(t)} + \sum_{i\in[N]} p_i(\Delta \mathtt{W}_{i,r_i}^{(t,E)} - \Delta \mathtt{W}_{i,r_i}^{(t,0)})$. Let $q_{\min} \triangleq \min_{l\in[L]} \sum_{i:r_{i,l}=r_l} p_i$, the sketching floor $S_1^{(t+1)} \triangleq \chi \cdot \mathbb{E}_t \left\| \mathcal{E}_{\text{sketch}}^{(t+1)} \right\|_F + 2\mu \mathbb{E}_t \left\| \mathcal{E}_{\text{sketch}}^{(t+1)} \right\|_F^2$, and $\overline{S_1} \triangleq \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[S_1^{(t+1)}\right]$. Theorem 1 shows that SeFoRA converges to a neighborhood of a first-order stationary point under Assumptions 1 to 4.

Theorem 1. *Suppose Assumptions 1 to 4 hold. Let the global minimum of the unregularized loss be $\mathcal{L}^{full,\star} \triangleq \min_{\Delta \mathtt{W}} \mathcal{L}^{full}(\Delta \mathtt{W})$ and $\Delta_0 \triangleq \mathbb{E}\left[\mathcal{L}^{full}(\Delta \mathtt{W}^{(0)})\right] - \mathcal{L}^{full,\star}$ and take a constant learning rate $\eta^{(t)} = \eta$. Let $\sigma_{l,j}^{(t)}$ be the j-th singular value of $\Delta \mathbf{W}_l^{(t)}$.* SeFoRA *satisfies:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(\mathbb{E}\left[\left\|\nabla_{\mathtt{A}}\mathcal{L}^{lora}(\mathtt{B}^{(t)},\mathtt{A}^{(t)})\right\|_F^2\right] + \mathbb{E}\left[\left\|\nabla_{\mathtt{B}}\mathcal{L}^{lora}(\mathtt{B}^{(t)},\mathtt{A}^{(t)})\right\|_F^2\right]\right) \le \frac{2\Delta_0}{\eta E T q_{\min}} + \underbrace{\frac{2\overline{S_1}}{\eta E q_{\min}}}_{\textit{sketching error}}$$

$$+\frac{2C_2\eta E}{q_{\min}} + \frac{2C_3\eta^2E^2}{q_{\min}^2} + \frac{4C_4\eta^3E^3}{q_{\min}} + \underbrace{\frac{6C_1}{2q_{\min}^2}\frac{\mu^2}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\sum_{i\in[N]} p_i \sum_{l\in[L]}\sum_{j=r_{i,l}+1}^{r_l}\left(\sigma_{l,j}^{(t)}\right)^2\Big]}_{\textit{clients local rank truncation}} + \underbrace{\frac{6C_1}{2q_{\min}^2}\zeta^2}_{\textit{client heterogeneity}}.$$

The proof is in Appendix E.1. Theorem 1 separates the optimization error from the three error sources: sketching, truncation to clients' local ranks, and client heterogeneity. The dependence on $q_{\min}$ indicates that the bound grows when clients with the maximum layer-wise rank have smaller aggregation weights. Under the spectral-tail condition in Assumption 5, we present Corollary 1 in which we select a constant learning rate that balances the optimization and sketching errors.

Assumption 5. *Let $\mathtt{V}^{(t+1)} \triangleq \Delta \mathtt{W}^{(t)} + \sum_{i\in[N]} p_i(\Delta \mathtt{W}_{i,r_i}^{(t,E)} - \Delta \mathtt{W}_{i,r_i}^{(t,0)})$. For all $l \in [L]$, there exist a fixed number $\varrho_l \in \{0, \cdots, r_l - 2\}$ and a constant $\overline{\tau_l} \ge 0$ such that, for every $T \ge 1$,*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\tau_{\varrho_l+1}^2(\mathbf{V}_l^{(t+1)})\right] \le \overline{\tau_l}^2.$$

Corollary 1. *Suppose Assumptions 1 to 5 hold, and $\Delta_0 > 0$. Define $c_l \triangleq (1+\frac{r_l}{k_l - r_l - 1})\cdot(1+\frac{\varrho_l}{r_l - \varrho_l - 1})$ and $\epsilon^2 \triangleq \sum_{l=1}^L c_l \cdot \overline{\tau_l}^2$. Choose $\gamma_T \triangleq \sqrt{\frac{\frac{\Delta_0}{T} + \chi\epsilon + 2\mu\epsilon^2}{C_2}}$ and $\eta = \frac{\gamma_T}{E}$,* SeFoRA *satisfies:*

$$\frac{1}{T}\sum_{t=0}^{T-1}\left(\mathbb{E}\left[\left\|\nabla_{\mathtt{A}}\mathcal{L}^{lora}(\mathtt{B}^{(t)},\mathtt{A}^{(t)})\right\|_F^2\right] + \mathbb{E}\left[\left\|\nabla_{\mathtt{B}}\mathcal{L}^{lora}(\mathtt{B}^{(t)},\mathtt{A}^{(t)})\right\|_F^2\right]\right)$$
$$\le \frac{4C_2}{q_{\min}}\gamma_T + \frac{2C_3}{q_{\min}^2}\gamma_T^2 + \frac{4C_4}{q_{\min}}\gamma_T^3 + \frac{6C_1}{2q_{\min}^2}\left(\mu^2\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\sum_{i\in[N]} p_i \sum_{l\in[L]}\sum_{j=r_{i,l}+1}^{r_l}\left(\sigma_{l,j}^{(t)}\right)^2\Big] + \zeta^2\right).$$

The proof is in Appendix E.2. Corollary 1 shows that SeFoRA converges with the rate $\mathcal{O}(T^{-1/2})$ through γ_T when the energy tail is bounded. Bounding the energy tail amounts to assuming a good sketching approximation, which we expect to hold in regimes where $\Delta \mathbf{W}$ is indeed of low rank.

5 Experiments

We conduct experiments across language understanding, mathematical reasoning, and image classification tasks, to evaluate the performance of SeFoRA in the rank-heterogeneous setting.

Models and datasets For language understanding, we fine-tune RoBERTa-Large (Liu et al., 2019) on SST-2, QNLI, QQP, and MNLI from the General Language Understanding Evaluation (GLUE) benchmark (Wang et al., 2018). For mathematical reasoning, we focus on GSM8K (Cobbe et al., 2021) using Llama 3.2 3B (Grattafiori et al., 2024) and Qwen 2.5 3B (Yang et al., 2024). For image classification, we use ViT (Dosovitskiy et al., 2020) model on CIFAR-100 (Krizhevsky et al., 2009), Tiny ImageNet (Le et al., 2015), Food-101 (Bossard et al., 2014), and DomainNet (Peng et al., 2019). We use 40 clients for GLUE, 20 for GSM8K, and 30 for the vision tasks, with full client participation. We show the stability of SeFoRA under the client's partial participation in Figure 3 of Appendix C.

Data heterogeneity For GLUE and the image datasets except DomainNet, we simulate data heterogeneity by partitioning the training data class-wise according to a Dirichlet distribution. The parameter α controls the level of heterogeneity, with smaller values corresponding to higher heterogeneity. We use $\alpha = 0.1$ unless otherwise specified. Since GSM8K has no class labels, we instead use a Dirichlet distribution to vary the number of training samples assigned to each client.

DomainNet contains images from 6 distinct domains, i.e., clipart, infograph, painting, quickdraw, real, and sketch, providing a naturally heterogeneous setting. We divide the 30 clients into 6 groups of five clients, with each group containing only images from a particular domain. Within each group, the data is distributed according to the Dirichlet distribution with parameters $\alpha = 0.1$ and 0.5. Consequently, DomainNet combines inter-group domain heterogeneity and intra-group label heterogeneity.

Rank heterogeneity We introduce rank heterogeneity across both clients and layers to simulate heterogeneous computation and memory budgets. The layer-wise rank profile assigns smaller ranks to earlier layers and gradually increases the rank towards the maximum in later layers. To model client-wise resource heterogeneity, we divide the clients into several tiers, each obtained by scaling the layer-wise rank profile by a different factor. In particular, for the GLUE experiments, we set $r_{\max} = 16$, and use the three scaling factors $0.5, 0.75$, and 1.0. For DomainNet, $r_{\max} = 16$, a five-tier profile is used, with scaling factors of $0.5, 0.625, 0.75, 0.875$, and 1.0, ensuring that each domain contains one client from each tier. The extreme vision experiment uses three equally sized client groups with $r_{\max} = 256$ and scaling factors $\frac{1}{32}$, $\frac{1}{16}$, and 1.0. In SeFoRA, we choose the sketch parameter $k_l = r_l + \max\{2, \lceil \frac{r_l}{8} \rceil\}$. The detailed configuration is reported in Figure 1 in Appendix B.

Baselines We compare SeFoRA with the three most relevant approaches: FlexLoRA, FSLoRA, and FedHL. FlexLoRA reconstructs and aggregates the client's full update from the LoRA adapters, thereby eliminating the bilinear mismatch at the cost of dense reconstruction at the server side. FSLoRA communicates selected components of fixed-rank local adapters, allowing clients to use heterogeneous local ranks without reconstructing dense updates; however, its aggregation does not eliminate the bilinear mismatch. FedHL employs error compensation together with dynamically adjusted aggregation weights to mitigate truncation-induced biases.

The implementation details are provided in Appendix B. We repeat each experiment using three random seeds, and report the mean accuracy and standard deviation. The best and second-best results are shown in bold and underline, respectively. Appendix C contains additional experiments, ablation studies, and the time spent on selected experiments.

5.1 Experiment Results

Domain and label heterogeneity We first evaluate the methods on DomainNet, which presents both domain and label heterogeneity. We report the results in Table 1. SeFoRA achieves the highest average accuracy in both settings, achieving results competitive with FSLoRA and consistently outperforming FlexLoRA and FedHL. These results demonstrate that SeFoRA outperforms the other methods when rank heterogeneity is combined with structured domain and label heterogeneity.

Mathematical reasoning and language understanding We report the accuracy on GSM8K using Llama 3.2 3B and Qwen 2.5 3B in Table 2. On Llama, SeFoRA achieves 36.2%, closely matching the 36.3% obtained by FSLoRA, and outperforming FlexLoRA and FedHL. On Qwen, SeFoRA achieves the highest accuracy of 56.1%. The results across the two models show that SeFoRA remains competitive with the strongest baselines on large language models. The GLUE results

Table 1: Accuracy (%) on DomainNet using ViT, with $r_{\max} = 16$.

	methods	clipart	infograph	painting	quickdraw	real	sketch	average
$\alpha = 0.1$	FlexLoRA	63.0 ± 0.5	31.3 ± 0.4	61.1 ± 0.3	32.6 ± 0.2	78.2 ± 0.2	53.9 ± 0.5	53.4 ± 0.3
	FSLoRA	66.7 ± 0.2	$\mathbf{32.9 \pm 0.4}$	63.4 ± 0.5	38.3 ± 0.4	$\mathbf{80.4 \pm 0.3}$	$\mathbf{56.9 \pm 0.1}$	56.4 ± 0.3
	FedHL	52.1 ± 1.0	25.3 ± 0.4	49.6 ± 1.0	21.4 ± 1.0	59.9 ± 0.3	44.3 ± 0.9	42.1 ± 0.7
	SeFoRA	$\mathbf{67.0 \pm 0.2}$	32.8 ± 0.2	$\mathbf{63.6 \pm 0.4}$	$\mathbf{38.9 \pm 0.2}$	80.2 ± 0.3	56.8 ± 0.1	$\mathbf{56.6 \pm 0.1}$
$\alpha = 0.5$	FlexLoRA	63.4 ± 0.2	32.0 ± 0.3	61.7 ± 0.4	33.5 ± 0.1	78.6 ± 0.3	54.3 ± 0.3	53.9 ± 0.1
	FSLoRA	67.5 ± 0.3	33.4 ± 0.3	$\mathbf{64.1 \pm 0.2}$	39.2 ± 0.4	80.8 ± 0.2	57.5 ± 0.4	57.1 ± 0.2
	FedHL	64.0 ± 0.3	31.1 ± 0.6	59.2 ± 1.3	27.5 ± 0.6	72.0 ± 1.6	53.3 ± 1.4	51.2 ± 0.9
	SeFoRA	$\mathbf{67.6 \pm 0.3}$	$\mathbf{33.8 \pm 0.1}$	64.0 ± 0.3	$\mathbf{41.2 \pm 0.3}$	$\mathbf{80.8 \pm 0.1}$	$\mathbf{57.7 \pm 0.1}$	$\mathbf{57.5 \pm 0.1}$

are reported in Table 2, showing that SeFoRA remains competitive with FSLoRA across the four language understanding tasks while avoiding aggregation with bilinear mismatch.

Table 2: Accuracy (%) on GSM8K and the GLUE tasks, all with $r_{\max} = 16$.

	Llama	Qwen	SST-2	QNLI	QQP	MNLI
FlexLoRA	34.5 ± 0.8	56.0 ± 0.4	$\mathbf{95.3 \pm 0.3}$	90.4 ± 1.3	70.6 ± 12.4	85.1 ± 2.5
FSLoRA	$\mathbf{36.3 \pm 1.1}$	55.7 ± 0.8	95.0 ± 0.1	$\mathbf{90.9 \pm 1.1}$	70.9 ± 13.5	$\mathbf{87.0 \pm 1.0}$
FedHL	21.5 ± 4.6	55.5 ± 0.7	93.8 ± 1.6	82.2 ± 5.9	$\mathbf{75.4 \pm 7.7}$	77.0 ± 10.6
SeFoRA	36.2 ± 0.2	$\mathbf{56.1 \pm 0.3}$	94.5 ± 1.0	90.7 ± 1.7	74.8 ± 10.5	84.3 ± 2.8

Extreme rank heterogeneity We consider a more extreme rank-heterogeneous setting, in which 30 clients are divided equally with maximum ranks $8, 16$ and 256. The rank distribution is illustrated in Figure 1. As shown in Table 3, SeFoRA consistently outperforms FSLoRA, with improvements of 0.8%, 1.0% and 0.7% on CIFAR-100, Tiny ImageNet, and Food-101. In comparison, under the standard rank-heterogeneous profile reported in Table 6 of Appendix C, the two methods differ by at most 0.3%. This shows the advantage of SeFoRA under extreme rank heterogeneity.

FlexLoRA achieves the highest accuracy; however, the federator reconstructs the dense update $\Delta\mathbf{W}_i$ for $i \in [N]$, incurring higher computational cost. In contrast, SeFoRA performs linear aggregation in the sketch space and approaches the accuracy of FlexLoRA without computing the dense updates $\Delta\mathbf{W}$. The fixed-size sketches may incur higher uplink communication costs for clients with small ranks. To reduce this overhead, SeFoRA can employ a flexible upload scheme in which client i transmits either $\mathbf{Y}_i$ and $\mathbf{Z}_i$ or $\mathbf{B}_i$ and $\mathbf{A}_i$, whichever is smaller. If the client uploads the LoRA factors, the server constructs the corresponding sketches. In this way, the communication for low-rank clients remains small while retaining sketch-space aggregation.

Table 3: Accuracy (%) on CIFAR-100, Tiny ImageNet, Food-101 under extreme rank heterogeneity.

	CIFAR-100	Tiny ImageNet	Food-101
FlexLoRA	$\mathbf{91.0 \pm 0.3}$	$\mathbf{88.0 \pm 0.2}$	$\mathbf{84.3 \pm 0.5}$
FSLoRA	88.9 ± 0.3	86.3 ± 0.2	83.3 ± 0.9
FedHL	85.0 ± 0.6	82.2 ± 0.3	76.8 ± 3.2
SeFoRA	89.7 ± 0.7	87.3 ± 0.3	84.0 ± 1.1

6 Conclusion

We propose SeFoRA, a federated LoRA method that addresses the rank heterogeneity challenge without facing the bilinear mismatch problem nor aggregating in the full ds-dimensional model space. SeFoRA constructs fixed-dimensional linear sketches and aggregates client updates in a common sketch space. Numerical experiments show that SeFoRA performs effectively across various rank- and data-heterogeneous settings and achieves competitive performance relative to the baselines.

The limitations of SeFoRA are a sketching error and increased communication costs for clients with lower ranks. The communication overhead for those clients can be mitigated by transmitting the adapters and requiring the federator to compute the corresponding sketches. Future work can study novel sketching methods that balance communication costs for clients without additional computation at the federator. Furthermore, the combination of differential privacy with SeFoRA remains an interesting future work that we will explore.

AI USE STATEMENT

We used ChatGPT to assist with debugging and verifying the software implementation of the proposed methods, screening mathematical proofs for potential logical gaps and inconsistencies, and generating the README.md file for the code. We used ChatGPT and Grammarly to help improve the language and grammar of the writing. We have not used generative AI tools for generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, generating the implementation, assisting with translation, cleaning and reformating dataset, supporting qualitative and thematic data analysis, and interpreting results. All AI-assisted changes were reviewed and validated by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All assumptions are clearly stated in the main text, and a detailed proof for every theoretical claim is provided in the appendix for checking and reproducing our results. The numerical experiments are explained in detail in the main body and the appendix, enabling the reader to replicate our experimental setup. In addition, we provide the source code as supplementary material and upload it to https://anonymous.4open.science/r/SeFoRA-7DAF. For every new idea, we explain it in text and summarize it in an Algorithm environment in the appendix for clarity of presentation.

REFERENCES

Sara Babakniya, Ahmed Roushdy Elkordy, Yahya H. Ezzeldin, Qingfeng Liu, Kee-Bong Song, Mostafa El-Khamy, and Salman Avestimehr. SLoRA: Federated parameter efficient fine-tuning of language models. In *Proc. NeurIPS Workshop Federated Learning Age Foundation Models (FL@FM)*, December 2023.

Jiamu Bai, Daoyuan Chen, Bingchen Qian, Liuyi Yao, and Yaliang Li. Federated fine-tuning of large language models under heterogeneous tasks and client resources. *Advances in Neural Information Processing Systems*, 37:14457–14483, 2024.

Jieming Bian, Lei Wang, Letian Zhang, and Jie Xu. LoRA-FAIR: Federated LoRA fine-tuning with aggregation and initialization refinement. *arXiv preprint arXiv:2411.14961*, 2024.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

Yuji Byun and Jaeho Lee. Towards federated low-rank adaptation of language models with rank heterogeneity. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 356–362, 2025.

Dongqi Cai, Yaozong Wu, Shangguang Wang, Felix Xiaozhu Lin, and Mengwei Xu. Efficient federated learning for modern NLP. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking (MobiCom)*, pp. 1–16, 2023.

Pushpita Chatterjee, Debashis Das, and Danda B. Rawat. Federated learning empowered recommendation model for financial consumer services. *IEEE Transactions on Consumer Electronics*, 2023.

Shuangyi Chen, Yue Ju, Hardik Dalal, Zhongwen Zhu, and Ashish Khisti. Robust federated finetuning of foundation models via alternating minimization of lora. *arXiv preprint arXiv:2409.02346*, 2024a.

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

Songlin Chen, Weicheng Wang, Xiaoliang Chen, Peng Lu, Zaiyan Yang, and Yajun Du. LLaMA-LoRA neural prompt engineering: A deep tuning framework for automatically generating chinese text logical reasoning thinking chains. *Data Intelligence*, 2024b.

Yae Jee Cho, Luyang Liu, Zheng Xu, Aldi Fahrezi, and Gauri Joshi. Heterogeneous LoRA for federated fine-tuning of on-device foundation models. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 12903–12913, 2024.

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.

Ning Ding, Xingtai Lv, Qiaosen Wang, Yulin Chen, Bowen Zhou, Zhiyuan Liu, and Maosong Sun. Sparse low-rank adaptation of pre-trained language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 4133–4145, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Tao Fan, Yan Kang, Guoqiang Ma, Weijing Chen, Wenbin Wei, Lixin Fan, and Qiang Yang. FATE-LLM: An industrial grade federated learning framework for large language models. *arXiv preprint arXiv:2310.10049*, 2023.

Wenzhi Fang, Dong-Jun Han, Liangqi Yuan, Seyyedali Hosseinalipour, and Christopher G Brinton. Federated sketching LoRA: On-device collaborative fine-tuning of large language models. In *International Conference on Machine Learning (ICML)*, 2026.

Xiuwen Fang and Mang Ye. Towards robust parameter-efficient fine-tuning for federated learning. *Advances in Neural Information Processing Systems*, 38:141777–141800, 2026.

Chun-Mei Feng, Yunlu Yan, Shanshan Wang, Yong Xu, Ling Shao, and Huazhu Fu. Specificity-preserving federated learning for mr image reconstruction. *IEEE Transactions on Medical Imaging*, 42(7):2010–2021, 2022.

Chun-Mei Feng, Bangjun Li, Xinxing Xu, Yong Liu, Huazhu Fu, and Wangmeng Zuo. Learning federated visual prompt in null space for mri reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8064–8073, 2023.

Sajjad Ghiasvand, Yifan Yang, Zhiyu Xue, Mahnoosh Alizadeh, Zheng Zhang, and Ramtin Pedarsani. Communication-efficient and tensorized federated fine-tuning of large language models. *arXiv preprint arXiv:2410.13097*, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Pengxin Guo, Shuang Zeng, Yanran Wang, Huijie Fan, Feifei Wang, and Liangqiong Qu. Selective aggregation for low-rank adaptation in federated learning. In *13th International Conference on Learning Representations Iclr 2025*, 2025.

Sihyeon Ha, Yongjeong Oh, and Yo-Seb Jeon. RB-LoRA: Rank-balanced aggregation for low-rank adaptation with federated fine-tuning. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 1737–1746, 2026.

Zeyu Han, Chao Gao, and Jinyang Liu. Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey. *Transactions on Machine Learning Research*, October 2024.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. LoRA+: Efficient low rank adaptation of large models. In *International Conference on Machine Learning (ICML)*, 2024.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Towards a Unified View of Parameter-Efficient Transfer Learning. In *International Conference on Learning Representations*, October 2021.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-Efficient Transfer Learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2790–2799. PMLR, May 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Meirui Jiang, Holger R. Roth, Wenqi Li, Dong Yang, Can Zhao, Vishwesh Nath, Daguang Xu, Qi Dou, and Ziyue Xu. Fair federated medical image segmentation via client contribution estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16302–16311, 2023.

Wenhao Jiang, Yuchuan Luo, Guilin Deng, Silong Chen, Xu Yang, Shihong Wu, Xinwen Gao, Lin Liu, and Shaojing Fu. Federated large language models: Feasibility, robustness, security and future directions. *arXiv preprint arXiv:2505.08830*, 2025.

Peter Kairouz and H Brendan McMahan. Advances and open problems in federated learning. *Foundations and trends in machine learning*, 14(1-2):1–210, 2021.

Junsu Kim, Jaeyeon Kim, and Ernest K Ryu. LoRA training provably converges to a low-rank global minimum or it fails loudly (but it probably won't fail). In *International Conference on Machine Learning*, volume 2025, 2025.

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

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

Weirui Kuang, Bingchen Qian, Zitao Li, Daoyuan Chen, Dawei Gao, Xuchen Pan, Yuexiang Xie, Yaliang Li, Bolin Ding, and Jingren Zhou. FederatedScope-LLM: A comprehensive package for fine-tuning large language models in federated learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5260–5271, 2024.

Yann Le, Xuan Yang, et al. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Seanie Lee, Sangwoo Park, Dong Bok Lee, Dominik Wagner, Haebin Seong, Tobias Bocklet, Juho Lee, and Sung Ju Hwang. Fedsvd: Adaptive orthogonalization for private federated learning with lora. *Advances in Neural Information Processing Systems*, 38:119733–119757, 2026.

Brian Lester, Rami Al-Rfou, and Noah Constant. The Power of Scale for Parameter-Efficient Prompt Tuning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.243.

Xiang Lisa Li and Percy Liang. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4582–4597, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.353.

Baohao Liao, Yan Meng, and Christof Monz. Parameter-Efficient Fine-Tuning without Introducing New Latency. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4242–4260, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.233.

Jin Liu, Yinbin Miao, Ning Xi, and Junkang Liu. Rethinking loRA for privacy-preserving federated learning in large models. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=BPzSV4uw0x.

Junkang Liu, Yuanyuan Liu, Fanhua Shang, Hongying Liu, Jin Liu, and Wei Feng. Improving generalization in federated learning with highly heterogeneous data via momentum-based stochastic controlled weight averaging. In *International Conference on Machine Learning (ICML)*, 2024a.

Junkang Liu, Fanhua Shang, Yuanyuan Liu, Hongying Liu, Yuangang Li, and YunXiang Gong. FedBCGD: Communication-efficient accelerated block coordinate gradient descent for federated learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 2955–2963, 2024b.

Junkang Liu, Fanhua Shang, Yuxuan Tian, Hongying Liu, and Yuanyuan Liu. Consistency of local and global flatness for federated learning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 3875–3883, 2025a.

Junkang Liu, Fanhua Shang, Junchao Zhou, Hongying Liu, Yuanyuan Liu, and Jin Liu. FedMuon: Accelerating federated learning with matrix orthogonalization. *arXiv preprint arXiv:2510.27403*, 2025b.

Junkang Liu, Fanhua Shang, Kewen Zhu, Hongying Liu, Yuanyuan Liu, and Jin Liu. FedAdamW: A communication-efficient optimizer with convergence and generalization guarantees for federated large models. *arXiv preprint arXiv:2510.27486*, 2025c.

Junkang Liu, Yuxuan Tian, Fanhua Shang, Yuanyuan Liu, Hongying Liu, Junchao Zhou, and Daorui Ding. DP-FedPGN: Finding global flat minima for differentially private federated learning via penalizing gradient norm. *arXiv preprint arXiv:2510.27504*, 2025d.

Muqing Liu, Chongjie Si, and Yuheng Jia. Flexlora: Entropy-guided flexible low-rank adaptation. *arXiv preprint arXiv:2601.22905*, 2026b.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-decomposed low-rank adaptation. In *International Conference on Machine Learning (ICML)*, 2024c.

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.

Guodong Long, Yue Tan, Jing Jiang, and Chengqi Zhang. Federated learning for open banking. In *Federated Learning: Privacy and Incentive*, pp. 240–254. Springer, 2020.

Yao Lu, Jiaxing QI, Zhongzhi Luan, Shaohan Huang, Carol Fung, Hailong Yang, and Depei Qian. FDLoRA: Personalized federated learning of large language model via dual lora tuning. *arXiv preprint arXiv:2406.07925*, 2026. URL https://arxiv.org/abs/2406.07925.

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

Wanyi Ning, Jingyu Wang, Qi Qi, Haifeng Sun, Daixuan Cheng, Cong Liu, Lei Zhang, Zirui Zhuang, and Jianxin Liao. Federated fine-tuning on heterogeneous loras with error-compensated aggregation. *IEEE Transactions on Neural Networks and Learning Systems*, 36(10):17826–17840, 2025.

Hongyi Peng, Han Yu, Xiaoxiao Li, and Qiang Yang. Rethinking LoRA for data heterogeneous federated learning: Subspace and state alignment. *arXiv preprint arXiv:2602.01746*, 2026.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1406–1415, 2019.

Zihao Peng, Jiandian Zeng, Boyuan Li, Guo Li, Shengbo Chen, and Tian Wang. Fedhl: Federated learning for heterogeneous low-rank adaptation via unbiased aggregation. *arXiv preprint arXiv:2505.18494*, 2025.

Chen Qiu, Xingyu Li, Chaithanya Kumar Mummadi, Madan Ravi Ganesh, Zhenzhen Li, Lu Peng, and Wan-Yi Lin. Text-driven prompt generation for vision-language models in federated learning. *arXiv preprint arXiv:2310.06123*, 2023.

Youyang Qu. Federated learning driven large language models for swarm intelligence: A survey. *arXiv preprint arXiv:2406.09831*, 2024.

Neranjan Senarath, Rohit Muralitharan, and Sadia Asif. Subspace-constrained federated learning with low-rank adaptation. *arXiv preprint arXiv:2606.22724*, 2026.

Raghav Singhal, Kaustubh Ponkshe, and Praneeth Vepakomma. FedEx-LoRA: Exact aggregation for federated and efficient fine-tuning of large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1316–1336. Association for Computational Linguistics, July 2025. ISBN 979-8-89176-251-0.

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

Chunlin Tian, Zhan Shi, Zhijiang Guo, Li Li, and Chengzhong Xu. HydraLoRA: An Asymmetric LoRA architecture for efficient fine-tuning. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, volume 37 of *NIPS '24*, pp. 9565–9584, Red Hook, NY, USA, June 2025. Curran Associates Inc.

Joel A Tropp, Alp Yurtsever, Madeleine Udell, and Volkan Cevher. Practical sketching algorithms for low-rank matrix approximation. *SIAM Journal on Matrix Analysis and Applications*, 38(4): 1454–1485, 2017.

Mojtaba Valipour, Mehdi Rezagholizadeh, Ivan Kobyzev, and Ali Ghodsi. DyLoRA: Parameter-efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3274–3287, 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. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pp. 353–355, 2018.

Ziyao Wang, Zheyu Shen, Yexiao He, Guoheng Sun, Hongyi Wang, Lingjuan Lyu, and Ang Li. FLoRA: Federated fine-tuning large language models with heterogeneous low-rank adaptations. *Advances in Neural Information Processing Systems*, 37:22513–22533, 2024.

Ming Wen, Jiaqi Zhu, Yuedong Xu, Yipeng Zhou, and Dingding Han. Differentially private federated low rank adaptation beyond fixed-matrix. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=TecJ926Vgn.

Fei Wu, Jia Hu, Geyong Min, and Shiqiang Wang. Adaptive rank allocation for federated parameter-efficient fine-tuning of language models. *IEEE Transactions on Computers*, 2026a.

Fei Wu, Jia Hu, Geyong Min, and Shiqiang Wang. Preventing rank collapse in federated low-rank adaptation with client heterogeneity. *arXiv preprint arXiv:2602.13486*, 2026b.

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 2026*, pp. 5087–5098, 2026a.

Peishen Yan, Yang Hua, Hao Wang, Jiaru Zhang, Xiaoyu Wu, Tao Song, and Haibing Guan. FedMomentum: Preserving lora training momentum in federated fine-tuning. *arXiv preprint arXiv:2603.08014*, 2026b.

Yunlu Yan, Hong Wang, Yawen Huang, Nanjun He, Lei Zhu, Yong Xu, Yuexiang Li, and Yefeng Zheng. Cross-modal vertical federated learning for mri reconstruction. *IEEE Journal of Biomedical and Health Informatics*, 2024.

Yunlu Yan, Chun-Mei Feng, Wangmeng Zuo, Rick Siow Mong Goh, Yong Liu, and Lei Zhu. Federated residual low-rank adaptation of large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Rui Ye, Rui Ge, Xinyu Zhu, Jingyi Chai, Yaxin Du, Yang Liu, Yanfeng Wang, and Siheng Chen. FedLLM-Bench: Realistic benchmarks for federated learning of large language models. *Advances in Neural Information Processing Systems*, 37:111106–111130, 2024a.

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 the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 6137–6147, 2024b.

Sixing Yu, J. Pablo Muñoz, and Ali Jannesari. Bridging the gap between foundation models and heterogeneous federated learning. *arXiv preprint arXiv:2310.00247*, 2023.

Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. BitFit: Simple Parameter-efficient Fine-tuning for Transformer-based Masked Language-models, September 2022.

Haoran Zhang, Dongjun Kim, Seohyeon Cha, and Haris Vikalo. Fedrot-loRA: Mitigating rotational misalignment in federated loRA. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=2X8Qi3VdjA.

Jianyi Zhang, Saeed Vahidian, Martin Kuo, Chunyuan Li, Ruiyi Zhang, Tong Yu, Guoyin Wang, and Yiran Chen. Towards building the federated GPT: Federated instruction tuning. In *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 6915–6919. IEEE, 2024.

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

Yicheng Zhang, Zhen Qin, Zhaomin Wu, Jian Hou, and Shuiguang Deng. Personalized federated fine-tuning for LLMs via data-driven heterogeneous model architectures. In *Proceedings of the ACM Web Conference 2026*, pp. 5099–5110, 2026b.

Haodong Zhao, Wei Du, Fangqi Li, Peixuan Li, and Gongshen Liu. Fedprompt: Communication-efficient and privacy-preserving prompt tuning in federated learning. In *ICASSP*, pp. 1–5, 2023.

Hongyun Zhou, Xiangyu Lu, Wang Xu, Conghui Zhu, Tiejun Zhao, and Muyun Yang. LoRA-drop: Efficient LoRA parameter pruning based on output evaluation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025a.

Yajie Zhou, Xiaoyi Pang, and Zhibo Wang. AFLoRA: Adaptive federated fine-tuning of large language models with resource-aware low-rank adaption. *arXiv preprint arXiv:2505.24773*, 2025b.

APPENDIX

A Algorithms **16**

B Implementation Details **16**

C Additional Experiment Results **16**

D Related Work **21**

E Proofs **24**

E.1 Proof of Theorem 1 28

E.2 Proof of Corollary 1 35

A ALGORITHMS

The algorithmic representation of SeFoRA is summarized in Algorithm 1. The sketch-inversion algorithm is described in Algorithm 2.

B IMPLEMENTATION DETAILS

Model configuration For RoBERTa-Large, the LoRA adapters are inserted into the query, key, and value projection matrices of every self-attention layer for RoBERTa-Large. For Llama 3.2 3B, Qwen 2.5 3B, and ViT, the LoRA adapters are inserted into the query and value matrices. The pretrained backbone parameters are frozen during federated fine-tuning. For classification tasks, the adapters and the classification head are optimized, while only the adapters are optimized for GSM8K. The adapter contribution is scaled by 0.1.

Federated optimization Unless otherwise stated, all clients participate in every communication round, with weight $p_i = \frac{|\mathcal{D}_i|}{\sum_{j=1}^{N} |\mathcal{D}_j|}$. Each client performs $E = 5$ local optimization steps per communication round. At each local step, clients randomly sample a mini-batch from their local training dataset, with batch sizes specified in Table 4 and Table 5. We use Adam as the optimizer, and sweep the learning rate over the values reported in Table 4 and Table 5. The total number of communication rounds for each task is also reported in the tables. The clients' rank distribution is illustrated in Figure 1.

Computation resources Experiments were conducted on machines equipped with NVIDIA RTX PRO 5000 Blackwell, NVIDIA RTX PRO 4500 Blackwell, or NVIDIA GeForce RTX 4090 GPUs. Each experimental run used a single GPU, with the GPU model determined by resource availability.

C ADDITIONAL EXPERIMENT RESULTS

Standard rank heterogeneity Table 6 shows the image classification results under the standard three-tier rank profile, as illustrated in subplot (b) in Figure 1. SeFoRA obtains the highest accuracy on Tiny ImageNet, and comparable accuracies on CIFAR-100 and Food-101. Together with the extreme rank-heterogeneous results in Table 3, and the results with varying maximum adapter rank, shown in Figure 2, these results show that SeFoRA remains competitive under standard rank heterogeneity, while its advantage over FSLoRA becomes more significant when the clients' rank profiles differ substantially.

Algorithm 1 SEFORA

Require: The pretrained model $\mathbf{W}_0$. Total rounds T and local steps E. In total, N clients, each client i holding their local dataset $\mathcal{D}_i$. The aggregation weights $p_i > 0$, with $\sum_{i=1}^{N} p_i = 1$. The local rank profiles $\mathbf{r}_i = (r_{i,1}, \cdots, r_{i,L})$. For $l \in [L]$, set $r_l \triangleq \max_{i\in[N]} r_{i,l}$ and $k_l > r_l + 1$. Set $r_{\max} \triangleq \max_{l\in[L]} r_l$, and $k_{\max} \triangleq \max_{l\in[L]} k_l$. Initialize global matrices $\mathbf{B}_l^{(0)} \in \mathbb{R}^{d\times r_l}$ as a full zero matrix and $\mathbf{A}_l^{(0)} \in \mathbb{R}^{r_l\times s}$ as a random Gaussian matrix. $\mathtt{B}^{(0)} = (\mathbf{B}_1^{(0)}, \cdots, \mathbf{B}_L^{(0)})$ and $\mathtt{A}^{(0)} = (\mathbf{A}_1^{(0)}, \cdots, \mathbf{A}_L^{(0)})$. A function UNSKETCHING$(\mathbf{Y}, \mathbf{Z}, \mathbf{\Psi})$ that unsketch the sketch matrices (Algorithm 2).

for round $t = 0, 1, \cdots, T-1$ **do**
  Federator initializes $\mathbf{\Psi}^{(t)} \in \mathbb{R}^{k_{\max}\times d}$ and $\mathbf{\Omega}^{(t)} \in \mathbb{R}^{s\times r_{\max}}$ from a standard normal distribution, and send the shared seed that generates $\mathbf{\Psi}^{(t)}$ and $\mathbf{\Omega}^{(t)}$ to all clients.
  Federator extracts $\mathbf{\Omega}_l^{(t)} = \mathbf{\Omega}_{:,1:r_l}^{(t)} \in \mathbb{R}^{s\times r_l}$ and $\mathbf{\Psi}_l^{(t)} = \mathbf{\Psi}_{1:k_l,:}^{(t)} \in \mathbb{R}^{k_l\times d}$, and computes $\mathbf{Y}_{\text{glo},l}^{(t)} = \mathbf{B}_l^{(t)}(\mathbf{A}_l^{(t)}\mathbf{\Omega}_l^{(t)})$, and $\mathbf{Z}_{\text{glo},l}^{(t)} = (\mathbf{\Psi}_l^{(t)}\mathbf{B}_l^{(t)})\mathbf{A}_l^{(t)}$ for all $l \in [L]$.
  For all clients i and layer l, federator computes $\mathbf{B}_{i,l}^{(t)} = (\mathbf{B}_l^{(t)})_{:,1:r_{i,l}}$ and $\mathbf{A}_{i,l}^{(t)} = (\mathbf{A}_l^{(t)})_{1:r_{i,l},:}$, and broadcasts $\mathtt{B}_i^{(t)}$ and $\mathtt{A}_i^{(t)}$ to the client i.
  for all clients i **in parallel do**
    for all adapter layer $l \in [L]$ **do**
      Regenerate $\mathbf{\Omega}_l^{(t)} = \mathbf{\Omega}_{:,1:r_l}^{(t)} \in \mathbb{R}^{s\times r_l}$ and $\mathbf{\Psi}_l^{(t)} = \mathbf{\Psi}_{1:k_l,:}^{(t)} \in \mathbb{R}^{k_l\times d}$ from the shared seed.
      Receive $\mathbf{B}_{i,l}^{(t)}$ and $\mathbf{A}_{i,l}^{(t)}$.
    end for
    Using $\mathbf{B}_{i,l}^{(t,0)} = \mathbf{B}_{i,l}^{(t)}$ and $\mathbf{A}_{i,l}^{(t,0)} = \mathbf{A}_{i,l}^{(t)}$, perform E local steps to obtain $\mathbf{B}_{i,l}^{(t,E)}$ and $\mathbf{A}_{i,l}^{(t,E)}$ for all $l \in [L]$.
    for all adapter layer $l \in [L]$ **do**
      Compute the sketch matrices

$$\mathbf{Y}_{i,l}^{(t)} = \mathbf{B}_{i,l}^{(t,E)}(\mathbf{A}_{i,l}^{(t,E)}\mathbf{\Omega}_l^{(t)}) - \mathbf{B}_{i,l}^{(t,0)}(\mathbf{A}_{i,l}^{(t,0)}\mathbf{\Omega}_l^{(t)}) \in \mathbb{R}^{d\times r_l},$$
$$\mathbf{Z}_{i,l}^{(t)} = (\mathbf{\Psi}_l^{(t)}\mathbf{B}_{i,l}^{(t,E)})\mathbf{A}_{i,l}^{(t,E)} - (\mathbf{\Psi}_l^{(t)}\mathbf{B}_{i,l}^{(t,0)})\mathbf{A}_{i,l}^{(t,0)} \in \mathbb{R}^{k_l\times s}.$$

    end for
    Send $\mathtt{Y}_i^{(t)} = (\mathbf{Y}_{i,1}^{(t)}, \cdots, \mathbf{Y}_{i,L}^{(t)})$ and $\mathtt{Z}_i^{(t)} = (\mathbf{Z}_{i,1}^{(t)}, \cdots, \mathbf{Z}_{i,L}^{(t)})$ to federator.
  end for
  for all adapter layer $l \in [L]$ **do**
    Federator aggregates

$$\mathbf{Y}_l^{(t+1)} = \mathbf{Y}_{\text{glo},l}^{(t)} + \sum_{i\in[N]} p_i\mathbf{Y}_{i,l}^{(t)} \quad \text{and} \quad \mathbf{Z}_l^{(t+1)} = \mathbf{Z}_{\text{glo},l}^{(t)} + \sum_{i\in[N]} p_i\mathbf{Z}_{i,l}^{(t)}.$$

    $(\widehat{\mathbf{B}}_l^{(t+1)}, \widehat{\mathbf{A}}_l^{(t+1)}) \leftarrow$ UNSKETCHING$(\mathbf{Y}_l^{(t+1)}, \mathbf{Z}_l^{(t+1)}, \mathbf{\Psi}_l^{(t)})$
    Federator computes $(\mathbf{U}_l^{(t+1)}, \mathbf{\Sigma}_l^{(t+1)}, \mathbf{V}_l^{(t+1)}) \leftarrow \text{SVD}(\widehat{\mathbf{A}}_l^{(t+1)})$, and

$$\mathbf{B}_l^{(t+1)} = \widehat{\mathbf{B}}_l^{(t+1)}\mathbf{U}_l^{(t+1)}(\mathbf{\Sigma}_l^{(t+1)})^{\frac{1}{2}} \quad \text{and} \quad \mathbf{A}_l^{(t+1)} = (\mathbf{\Sigma}_l^{(t+1)})^{\frac{1}{2}}(\mathbf{V}_l^{(t+1)})^{\top}.$$

  end for
end for
Return $(\mathtt{B}^{(T)}, \mathtt{A}^{(T)})$

Algorithm 2 Sketch-inversion Algorithm

Require: Given the random matrix $\mathbf{\Psi} \in \mathbb{R}^{k\times d}$ drawn from a standard normal distribution, and two sketch matrices $\mathbf{Y} \in \mathbb{R}^{d\times r}$ and $\mathbf{Z} \in \mathbb{R}^{k\times s}$, produced from the target large matrix $\mathbf{W} \in \mathbb{R}^{d\times s}$, i.e., $\mathbf{Y} = \mathbf{W\Omega}$ and $\mathbf{Z} = \mathbf{\Psi W}$.
Ensure: Return the low-rank factors $\widehat{\mathbf{B}} \in \mathbb{R}^{d\times r}$ and $\widehat{\mathbf{A}} \in \mathbb{R}^{r\times s}$ of a rank-r approximation of $\mathbf{W}$.

function UNSKETCHING$(\mathbf{Y}, \mathbf{Z}, \mathbf{\Psi})$
  $(\widehat{\mathbf{B}}, \sim) \leftarrow \text{QR}(\mathbf{Y})$
  $(\mathbf{Q}, \mathbf{R}) \leftarrow \text{QR}(\mathbf{\Psi}\widehat{\mathbf{B}})$
  $\widehat{\mathbf{A}} = \mathbf{R}^{-1}(\mathbf{Q}^T\mathbf{Z})$
  Return $(\widehat{\mathbf{B}}, \widehat{\mathbf{A}})$
end function

Table 4: The detailed configuration for the language understanding and mathematical reasoning tasks.

	GLUE	GSM8K
Model	RoBERTa-Large	Llama 3.2 3B and Qwen 2.5 3B
Clients	40	20
Batch Size	64	4
Maximum length	128	512
Communication rounds	400	400 for Llama, 200 for Qwen
Local steps	5	5
Learning rates	$\{5e-4, 1e-3, 2e-3, 3e-3, 4e-3\}$	$\{1e-5, 2e-5, 3e-4, 7e-4, 1e-3, 2e-3\}$
Adapting layers	24	28 for Llama, 36 for Qwen
Metric	Accuracy	Exact-match accuracy

Table 5: The detailed configuration for the image classification tasks.

	CIFAR-100/Tiny ImageNet/Food-101/DomainNet
Model	ViT-Base
Clients	30
Batch Size	32
Communication rounds	300
Local steps	5
Learning rates	$\{1e-3, 3e-3, 5e-3, 7e-3, 9e-3, 1e-2, 2e-2\}$
Adapting layers	12
Metric	Accuracy

Table 6: Accuracy (%) on CIFAR-100, Tiny ImageNet, and Food-101 using ViT under standard rank heterogeneity, with $r_{\max} = 16$. For SeFoRA, we choose $k = r + \max\{2, \lceil r/8 \rceil\}$.

	CIFAR-100	Tiny ImageNet	Food-101
FlexLoRA	$\mathbf{91.1 \pm 0.2}$	87.9 ± 0.1	84.2 ± 0.6
FSLoRA	$\underline{91.0 \pm 0.1}$	$\underline{88.4 \pm 0.2}$	$\mathbf{86.2 \pm 0.8}$
FedHL	83.4 ± 1.2	79.5 ± 1.3	75.6 ± 2.9
SeFoRA	$\underline{91.0 \pm 0.2}$	$\mathbf{88.7 \pm 0.2}$	$\underline{86.1 \pm 0.3}$

Impact of rank We vary the maximum adapter rank over $\{2, 4, 8, 16, 32, 64\}$, and the same rank configuration is applied to all the methods at each value. Across all four methods, performance generally improves as the maximum rank $r_{\max}$ increases, whereas FedHL is more sensitive to the choice of r_{max}. These results indicate that the performance of SeFoRA is relatively less sensitive to the choice of the adapter rank, and remains competitive across the three datasets.

Impact of partial participation and the number of clients In Figure 3, we vary the clients' participation rate over $\{0.2, 0.5, 1\}$, while the total number of clients remains $n = 30$. The clients are sampled uniformly without replacement in each round. The results show that SeFoRA remains robust when not all clients participate in each communication round and remains competitive with other methods.

Figure 4 evaluates SeFoRA with $10, 30, 100$ and 200 clients, with the batch size and the number of local steps fixed, and all clients participate in every communication round. With the increase in the number of clients from 10 to 30, the accuracy on all three datasets improves. However, the results remain nearly unchanged as the number of clients continues to increase. Overall, SeFoRA's performance does not degrade as the size of the federated learning network increases; instead, it gradually approaches saturation as the number of participants grows.

Impact of sketch size We first vary the sketching dimension k from $k = r + 2$ to $k = 2r + 1$, while fixing the adapter rank to $r_{\max} = 16$ on the GLUE tasks, as shown in Figure 5. The performance on the four datasets remains largely unchanged for $k_{\max} \in \{18, 20, 24, 33\}$.

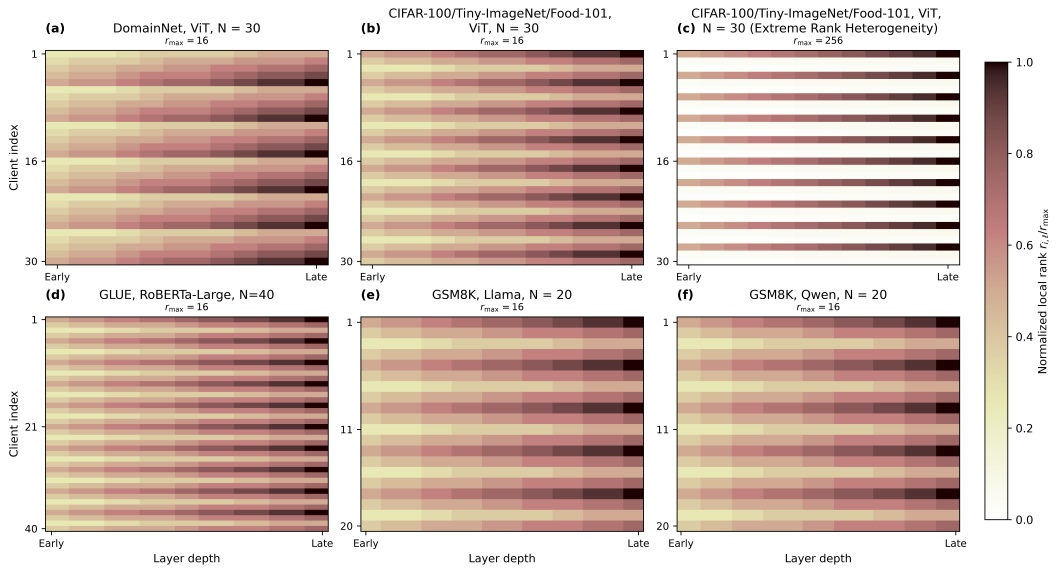

Figure 1: We report the client-wise and layer-wise heterogeneous rank profile used in the experiments. Rows are the clients, columns correspond to the adapted layers from early to late stage, and colors indicate ranks normalized by the maximum rank $r_{\max}$. Specifically, (a) is used in the experiment producing Table 1, (b) is used for Table 6 and Figure 3, Figure 4 and Figure 6, (c) is used for Table 3, (d), (e) and (f) are used to produce Table 2.

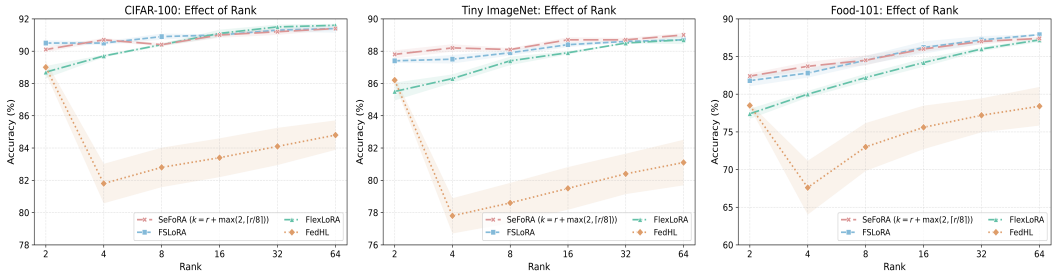

Figure 2: Impact of the maximum adapter rank $r_{\max} \in \{2, 4, 8, 16, 32, 64\}$ across all methods. For SeFoRA, we choose $k = r + \max\{2, \lceil r/8 \rceil\}$.

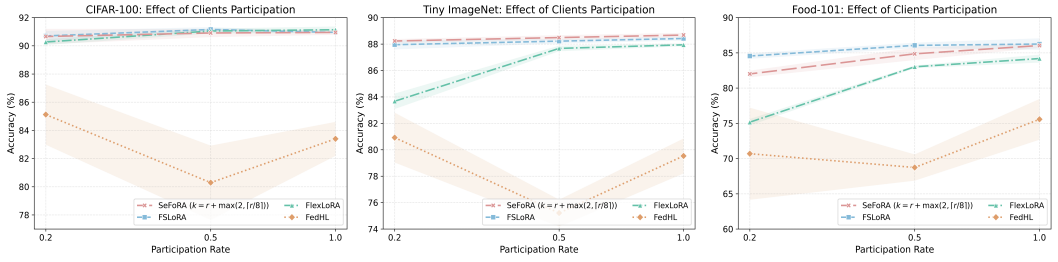

Figure 3: Impact of different clients' participation rates $\{0.2, 0.5, 1.0\}$.

We further compare different choices of k while varying the adapter rank $r_{\max}$, as illustrated in Figure 6. The results show that using the fixed sketching dimension $k = r + 2$ becomes insufficient at large ranks, e.g., $r_{\max} = 32, 64$, with the largest degradation observed on Food-101. In contrast, using a substantially larger sketch size, such as $k = 2r + 1$ or $k = r + \max\{2, \lceil \frac{r}{2} \rceil\}$, does not provide a consistent improvement over the smaller size $k = r + \max\{2, \lceil \frac{r}{8} \rceil\}$. The selected choice $k = r + \max\{2, \lceil \frac{r}{8} \rceil\}$ therefore provides a balance between the performance and the dimension of

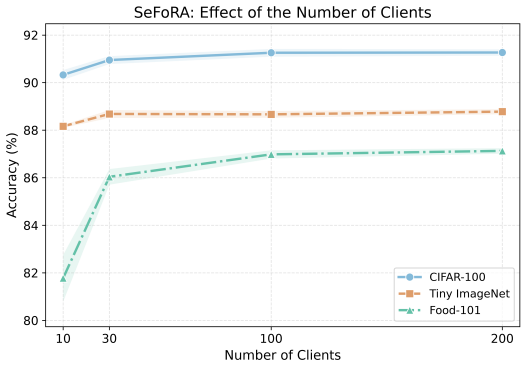

Figure 4: Impact of the number of clients $N \in \{10, 30, 100, 200\}$ on SeFoRA across the three image classification datasets. We choose $r_{\max} = 16$ and $k = r + \max\{2, \lceil r/8 \rceil\}$.

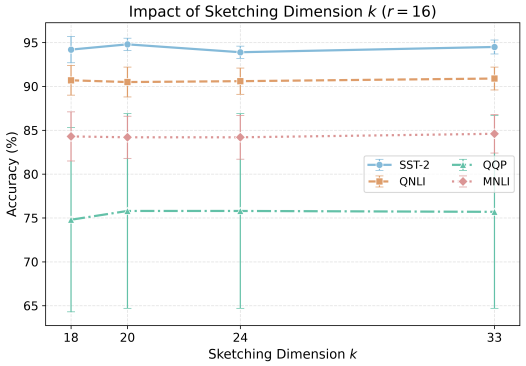

Figure 5: Impact of sketching dimension $k_{\max} \in \{18, 20, 24, 33\}$ on GLUE accuracy (%), with $r_{\max} = 16$.

the sketching. Note that, when $r \leq 4$, $k = r + \max\{2, \lceil \frac{r}{2} \rceil\}$, $r + \max\{2, \lceil \frac{r}{8} \rceil\}$ and $r + 2$ produce the same value of $k = r + 2$, and when $8 \leq r \leq 16$, $k = r + \max\{2, \lceil \frac{r}{8} \rceil\}$ and $k = r + 2$ also result in the same value of $k = r + 2$.

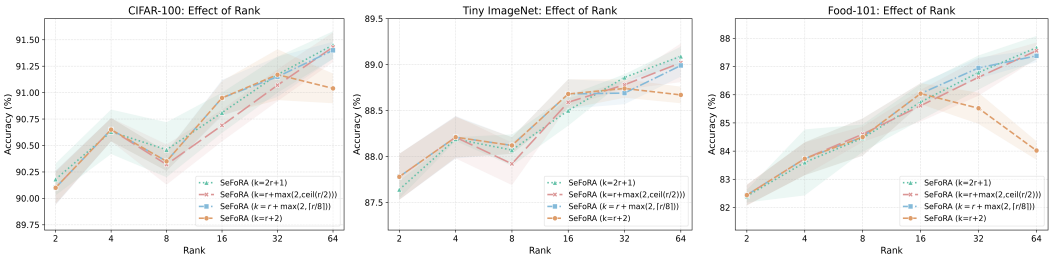

Figure 6: Impact of different sketching dimensions k on SeFoRA across maximum rank $r_{\max} \in \{2, 4, 8, 16, 32, 64\}$.

System Metrics We measure computational cost under a matched experimental configuration on QNLI using RoBERTa-Large, reported in Table 7. All methods are trained for 400 communication rounds, with 40 clients with full participation, a batch size of 64, five local steps, $r_{\max} = 16$, and standard rank heterogeneity on a single NVIDIA RTX PRO 4500 Blackwell GPU. SeFoRA requires 2.40 ± 0.15 GPU-hours per run, comparable to FSLoRA and lower than FedHL and FlexLoRA. We additionally report the mean server aggregation time per communication round, averaged across

	GPU hours per run (h)	Server aggregation time per round (s)
FlexLoRA	2.89	4.821
FSLoRA	**2.38**	<u>0.127</u>
FedHL	2.78	3.866
SeFoRA	<u>2.40</u>	**0.102**

Table 7: System efficiency on QNLI with RoBERTa-Large.

rounds and seeds. SeFoRA reduces server aggregation complexity compared with methods that require dense reconstruction and a full SVD, such as FlexLoRA and FedHL.

D Related Work

Parameter-efficient fine-tuning (PEFT). PEFT methods adapt a pretrained model $\mathbf{W}_0$ by updating only a small, structured subset of parameters. They fall into three families: *additive* (inserting trainable modules (Houlsby et al., 2019; He et al., 2021; Li & Liang, 2021; Lester et al., 2021)), *selective* (updating a subset of existing weights (Zaken et al., 2022; Liao et al., 2023)), and *reparametrized* (imposing low-dimensional structure on the update (Hu et al., 2022; Liu et al., 2024c; Tian et al., 2025; Zhang et al., 2023; Valipour et al., 2023)); see (Han et al., 2024) for a survey.

Among reparametrized methods, LoRA (Hu et al., 2022) is perhaps the most popular method due to its simplicity and negligible inference overhead. The weight updates are parametrized as $\Delta\mathbf{W} = \mathbf{BA}$ with *factor matrices* $\mathbf{B} \in \mathbb{R}^{d\times r}$, $\mathbf{A} \in \mathbb{R}^{r\times s}$, $r \ll \min(d, s)$, thus reducing the number of trainable parameters from ds to $(d + s)r$. Beyond the original LoRA formulation (Hu et al., 2022), numerous variants have been proposed to improve adaptation efficiency, convergence, and parameter utilization. LoRA+ (Hayou et al., 2024) accelerates optimization by assigning different learning rates to the two low-rank matrices. DoRA (Liu et al., 2024c) improves fine-tuning performance by decomposing pretrained weights into magnitude and directional components, allowing LoRA to optimize only the directional updates. Other methods focus on improving parameter efficiency through adaptive resource allocation. LoRA-drop (Zhou et al., 2025a) prunes redundant LoRA parameters based on their contribution to the model output, whereas AdaLoRA (Zhang et al., 2023) dynamically reallocates the adaptation rank across layers according to their relative importance under a fixed parameter budget. HydraLoRA (Tian et al., 2025) modifies the factorization structure. Furthermore, LoRA has been adopted in application-specific settings, such as automatic prompt engineering for logical reasoning (Chen et al., 2024b). On the theory side, (Kim et al., 2025) shows that ℓ_2-regularized LoRA is equivalent to nuclear-norm-regularized full fine-tuning subject to a rank constraint.

PEFT and federated learning Federated learning (FL) has become the de facto paradigm for distributed machine learning, enabling collaborative model training while keeping data decentralized. By allowing clients to train models locally without sharing raw data, FL offers better privacy guarantees and has consequently found widespread adoption in privacy-sensitive applications, including finance (Long et al., 2020; Chatterjee et al., 2023) and healthcare (Feng et al., 2022; Jiang et al., 2023; Feng et al., 2023; Yan et al., 2024). Motivated by the success of large language models (LLMs), recent research has extended FL to support their efficient fine-tuning and deployment (Qu, 2024; Fan et al., 2023; Zhang et al., 2024). Several frameworks have subsequently been proposed to accelerate research in this emerging area. FederatedScope-LLM (Kuang et al., 2024) provides a comprehensive platform for federated fine-tuning and instruction tuning of LLMs, while OpenFedLLM (Ye et al., 2024b) offers a unified framework for training LLMs over decentralized private data and systematically studies federated instruction tuning and value alignment. Complementing these systems, FedLLM-Bench (Ye et al., 2024a) introduces realistic benchmarks and extensive empirical evaluations for federated fine-tuning, enabling standardized assessment of optimization algorithms and communication-efficient methods. Together, these works have established the foundational infrastructure, benchmarks, and evaluation methodologies that drive the rapidly growing field of federated fine-tuning of LLMs.

Combining PEFT with FL enables for adapting large foundation models on vast amount of data while substantially reducing communication overhead, computational complexity, and privacy risks. Existing approaches span adapter-based methods (Cai et al., 2023; Ghiasvand et al., 2024), prompt

tuning (Zhao et al., 2023; Qiu et al., 2023), selective parameter optimization (Yu et al., 2023), and low-rank gradient subspace optimization (Peng et al., 2026). More recently, Low-Rank Adaptation (LoRA) (Hu et al., 2022) has become a popular PEFT technique on its own and in FL.

Federated LoRA. Early works such as FedIT (Zhang et al., 2024) integrate LoRA into the standard FedAvg framework for federated instruction tuning, while subsequent methods address the limitations of naive LoRA aggregation under heterogeneous data and system conditions. However, combining LoRA with FL raises several challenges that do not exist in centralized LoRA or in federated averaging alone. Namely: the error introduced by the bilinear mismatch resulting from first aggregating the $\mathbf{A}_i$'s and $\mathbf{B}_i$'s and then computing $\overline{\Delta \mathbf{W}} = (\sum_i p_i \mathbf{B}_i)(\sum_i p_i \mathbf{A}_i)$ as done in FedIT (Zhang et al., 2024); or the rank-r approximation error resulting from first computing $\Delta \mathbf{W} = \sum_i p_i \mathbf{B}_i \mathbf{A}_i$ which is not necessarily rank-r and then computing a rank-r approximation, e.g., using SVD decomposition as in FlexLoRA (Bai et al., 2024). In addition, the proposed aggregation rule should account for the *rank heterogeneity* problem: when clients use different ranks r_i, the factor matrices have mismatched dimensions, rendering direct factor averaging infeasible. To overcome those hurdles, several approaches have been proposed, as we detail in the sequel. Beyond model parameterization, several recent works have focused on improving the optimization process itself. FedBCGD (Liu et al., 2024b) introduces an accelerated block coordinate gradient descent algorithm, FedSWA (Liu et al., 2024a) employs stochastic weight averaging to improve generalization under highly non-IID data, FedAdamW (Liu et al., 2025c) proposes a communication-efficient AdamW-inspired optimizer for federated large models, FedNSAM (Liu et al., 2025a) analyzes the relationship between local and global sharpness to improve optimization consistency, FedMuon (Liu et al., 2025b) accelerates convergence through matrix orthogonalization, and DP-FedPGN (Liu et al., 2025d) encourages globally flatter minima in differentially private FL by penalizing gradient norms. Client data heterogeneity is tackled in (Yan et al., 2025). In LA-LoRA (Liu et al., 2026a), during local iterations, the clients alternately update $\mathbf{B}$ and $\mathbf{A}$ and send both updated adapters to the federator for aggregation. This alternation is shown to provide stability during training and to allow for differential privacy with lower noise. FedASK (Wen et al., 2025) uses a double-sketching approach primarily to provide a differential privacy guarantee, but does not account for rank-heterogeneity. Robustness to label noise in federated LoRA is studied in (Fang & Ye, 2026). Personalized federated adaptation is also considered, e.g., (Zhang et al., 2026b; Lu et al., 2026).

Tackling the bilinear mismatch The bilinear mismatch introduces an error of $O(\eta^2)$ per round for a single local step and grows as $O(\eta^2 E^2)$ for multiple local steps (Wang et al., 2024; Singhal et al., 2025) and is more prominent when considering differential privacy guarantees. In (Sun et al., 2024), FFA-LoRA freezes the initialized adapter $\mathbf{A}$ and trains only the adapter $\mathbf{B}$. As such, the federator avoids the bilinear mismatch problem. However, freezing the adapter $\mathbf{A}$ reduces the expressiveness of the adaptation and affects the performance. In FedSVD (Lee et al., 2026), the clients only train $\mathbf{B}$ locally, while the global adapter $\mathbf{A}^{(t)}$ is computed based on an SVD of the matrix $\mathbf{B}^{(t)}\mathbf{A}^{(t-1)}$; thus enabling the update of the factor $\mathbf{A}$ while clients still train and aggregate only $\mathbf{B}$. In a similar spirit, FedSA-LoRA (Guo et al., 2025) trains $\mathbf{A}$ globally while each client trains $\mathbf{B}$ locally. It provides a formal convergence analysis; however, since $\mathbf{B}$ is different for each client, the algorithm is suitable for personalized FL. The concurrent works introducing LoRA-A^2 (Koo et al., 2025) and RoLoRA (Chen et al., 2026) alternate between freezing $\mathbf{A}$ training $\mathbf{B}$, and training $\mathbf{A}$ freezing $\mathbf{B}$. In LoRA-A^2, the federator selects an adaptive rank for each client to handle data heterogeneity. The client updates only a subset of all possible ranks, using a mask to perform the update without modifying the dimensionality of the adapters; however, the paper lacks a theoretical foundation. In RoLoRA, all clients maintain the same rank, and a convergence guarantee is given. In (Zhang et al., 2026a), the authors propose FedRot-LoRA, which takes a different approach: aligning local subspaces via orthogonal rotations at the clients' side before aggregating. FedEx-LoRA (Singhal et al., 2025) keeps the LoRA factors trainable across rounds and absorbs the mismatch into the frozen pretrained weights as a residual term; however, it increases both computation and communication overheads and does not provide a convergence guarantee. This reduces the bilinear mismatch error but does not eliminate it. LoRA-FAIR (Bian et al., 2024) introduces a correction term in $\mathbf{B}$ to prevent the $\overline{\mathbf{B}}\,\overline{\mathbf{A}}$ from deviating too much from the actual update $\sum_i p_i \mathbf{B}_i \mathbf{A}_i$. Despite their advantages, those methods are not applicable in rank-heterogeneous settings.

Rank-heterogeneous federated LoRA In (Babakniya et al., 2023), the authors investigate methods for finding a good initialization of the adapters via full fine-tuning and then running federated LoRA, without altering the federated LoRA process. Adapting the federation of LoRA to account for rank heterogeneity is proving to be more successful. The authors of (Wang et al., 2024) propose FLoRA, in which the federator stacks the received $\mathbf{A}_i$'s and $\mathbf{B}_i$'s from the clients in a way that the multiplication of the stacked matrices emulates the computation and aggregation of the local $\Delta\mathbf{W}_i$'s. As such, the globally updated $\mathbf{A}$ and $\mathbf{B}$ have larger dimensions, and the resulting $\Delta\mathbf{W}$ has a rank that could be larger than r; thus increasing the communication and computation cost at the clients. Furthermore, in each round, the updated global model is integrated into the base model and the adapters are reinitialized, thereby potentially losing information about the adapters obtained from the previous rounds. Bai et al. (Bai et al., 2024) took another approach in their FlexLoRA. The federator computes and aggregates the clients' $\Delta\mathbf{W}_i$'s, and, for each client, the federator uses SVD to decompose and send a rank-r_i approximation of the global $\Delta\mathbf{W}$. This method incurs an $O(d^2s)$ computational cost at the server, and the paper analyzes the number of samples the clients need to have to obtain a good generalization bound, but it does not give a proof of convergence for the algorithm. Similarly, in FedMomentum (Yan et al., 2026b), the federator computes and aggregates the local $\Delta\mathbf{W}_i$'s to obtain $\Delta\mathbf{W}$ and then performs a few mathematical transformations on $\Delta\mathbf{W}$ to efficiently compute an approximation of its SVD decomposition. A key difference in this work is computing the global adapters as $\mathbf{B} = \widetilde{\mathbf{U}}\Sigma^{1/2}$ and $\mathbf{A} = \Sigma^{1/2}\mathbf{V}^\top$, where $\widetilde{\mathbf{U}}$, Σ and $\mathbf{V}$ are the matrices obtained from the approximation of the SVD of $\Delta\mathbf{W}$. The computation complexity at the federator is $O(dN^2r^2)$. Also, no convergence guarantee is given. Similarly, in FedSRD (Yan et al., 2026a) the clients apply local sparsification and pruning techniques to reduce the ranks of their local adapters, and the federator reconstructs the clients' full update $\mathbf{B}_i\mathbf{A}_i$ and aggregates in the full-rank space, which increases the computation at the server. The server then computes the SVD of $\Delta\mathbf{W}$ to compute a low-rank update and uses sparsification techniques to further reduce the communication cost; this work also lacks a convergence analysis. In HetLoRA (Cho et al., 2024), the clients prune the adapters to smaller ranks by adding a regularization factor to the optimization problem. Then, the federator zero-pads the clients' $\mathbf{A}_i$'s and $\mathbf{B}_i$'s to ensure they have the same dimension. However, when aggregating the zero-padded adapters, a careful weighted average is performed. For each adapter, a weight p_i proportional to the norm of the singular value vector of $\Delta\mathbf{W}_i$ is computed. The computational complexity of this method is $O(Ndrs)$ for computing the global update. Similarly, no convergence guarantee is given. FSLoRA (Fang et al., 2026) modifies the update to $\mathbf{BSA}$ to embed the sketching matrices $\S_i$ at the clients and account for rank heterogeneity. The matrices $\mathbf{S}_i$ are $r \times r$ diagonal matrices with only r_i non-zero entries chosen at random. The algorithm can be theoretically proven to converge with a rate of $O(\widetilde{L}/\sqrt{NT})$ when run for T iterations. However, the choice of which r_i entries are non-zero is random and not based on the properties of client's data. The authors of (Byun & Lee, 2025) replace the zero-padding by replicating rows of $\mathbf{A}$ and columns of $\mathbf{B}$ to make the matrices have the same ranks. In RB-LoRA (Ha et al., 2026), a unifying framework for completing the local adapters (via replication) and stacking them for computation at the federator is proposed. A client's local adapters are weighted based on the amount of data the client holds, proportional to the total number of data (for $\mathbf{A}$) and to the total number of data held by clients with ranks less than or equal to the client's (for $\mathbf{B}$). The method's performance is evaluated via numerical experiments. In (Zhou et al., 2025b), the proposed AFLoRA adds a diagonal matrix $\mathbf{\Lambda}$ of dimension $r \times r$ to the clients' local adapters to have $\Delta\mathbf{W}_i = \mathbf{B}_i\mathbf{\Lambda}_i\mathbf{A}_i$. The diagonal matrix is meant to compute the contribution of each dimension of $\mathbf{B}_i$ to the local $\Delta\mathbf{W}$. Then, the clients accordingly adjust their local ranks. The adapter $\mathbf{A}$ is only trained at the federator using a public dataset, and the $\mathbf{B}_i$ and $\mathbf{\Lambda}_i$ are trained at the clients. To aggregate the local adapters, zero padding is used, and aggregation weights proportional to the clients' data volume and local LoRA ranks are computed. FedARA (Wu et al., 2026a) adaptively allocates local ranks to each layer at clients, and layers with small ranks are not updated. Through a majority vote at the federator, global layers will also be pruned. However, aggregating client updates still requires zero-padding and matrices of the same dimension. Recent work has further examined the biases caused by heterogeneous ranks (Wu et al., 2026b; Peng et al., 2025). In (Wu et al., 2026b), rank collapse is identified, i.e., the energy of the aggregated update progressively concentrates on the smallest rank shared by the participating clients. They propose raFLoRA, which splits each local update into rank partitions and aggregates them separately. This prevents higher-rank components from being diluted by clients who do not update them. Nevertheless, the server reconstructs the partitioned updates in the full model space and applies an SVD to recover global factors, incurring computational overhead at the server. FedHL (Peng et al., 2025) focuses

on the truncation-induced bias arising in federated LoRA with heterogeneous ranks. The method maintains a full-rank global model as a calibrated aggregation basis, and the truncation bias can be eliminated by using a compensation mechanism. ECLoRA (Ning et al., 2025) reconstructs the full rank update at the server and utilizes a randomized SVD for broadcasting. It also introduces a compensation term to further improve the performance of the algorithm.

In (Senarath et al., 2026), the authors introduce a subspace regularization term in the objective, making sure the subspaces spanned by the clients' low-rank updates stay aligned. However, in each round, the clients send $\Delta\mathbf{W}_i = \mathbf{B}_i\mathbf{A}_i$ to the federator, increasing the communication overhead between the clients and the federator.

E PROOFS

In the proofs below, we adopt the following notation. Let $\mathcal{F}_t$ represent all algorithmic history up to the beginning of round t. We define $\mathbb{E}_t[\cdot] \triangleq \mathbb{E}[\cdot \mid \mathcal{F}_t]$ as the conditional expectation taking into account all sources of randomness occurring during round t. Let $\mathbb{E}_{\xi_i}[\cdot]$ denote the expectation taken with respect to the local mini-batch sampling at round t, conditioned on the model state at the moment of sampling.

For each client i, let $\mathbf{r}_i = (r_{i,1}, \cdots, r_{i,L})$ denote its rank profile. For each layer l, define the truncation operation $\mathfrak{d}(\cdot)$ through a matrix $\mathbf{\Gamma}_{i,l} = \begin{bmatrix} \mathbf{I}_{r_{i,l}} \\ \mathbf{0} \end{bmatrix} \in \mathbb{R}^{r_l \times r_{i,l}}$, and the truncation tuple is $\Gamma_i \triangleq \{\mathbf{\Gamma}_{i,1}, \cdots, \mathbf{\Gamma}_{i,L}\}$. For global factors $\mathtt{B}^{(t)}$ and $\mathtt{A}^{(t)}$, define $[\![\mathtt{B}^{(t)}]\!]_{r_i} \triangleq (\mathbf{B}_l\mathbf{\Gamma}_{i,l})_{l=1}^{L}$ and $[\![\mathtt{A}^{(t)}]\!]_{r_i} \triangleq ((\mathbf{\Gamma}_{i,l})^\top\mathbf{A}_l)_{l=1}^{L}$. We further define

$$
\begin{aligned}
g^{(t)}_{\mathtt{A},i,r_i} &= \nabla_{\mathtt{A}}\mathcal{L}_i^{\text{lora}}([\![\mathtt{B}^{(t)}]\!]_{r_i}, [\![\mathtt{A}^{(t)}]\!]_{r_i}), & g^{(t)}_{\mathtt{B},i,r_i} &= \nabla_{\mathtt{B}}\mathcal{L}_i^{\text{lora}}([\![\mathtt{B}^{(t)}]\!]_{r_i}, [\![\mathtt{A}^{(t)}]\!]_{r_i}) \\
\hat{g}^{(t,e)}_{\mathtt{A},i,r_i} &= \nabla_{\mathtt{A}}\mathcal{L}_i^{\text{lora}}(\mathtt{B}^{(t,e)}_{i,r_i}, \mathtt{A}^{(t,e)}_{i,r_i}; \xi_i^{(t,e)}), & \hat{g}^{(t,e)}_{\mathtt{B},i,r_i} &= \nabla_{\mathtt{B}}\mathcal{L}_i^{\text{lora}}(\mathtt{B}^{(t,e)}_{i,r_i}, \mathtt{A}^{(t,e)}_{i,r_i}; \xi_i^{(t,e)}) \\
g^{(t,e)}_{\mathtt{A},i,r_i} &= \nabla_{\mathtt{A}}\mathcal{L}_i^{\text{lora}}(\mathtt{B}^{(t,e)}_{i,r_i}, \mathtt{A}^{(t,e)}_{i,r_i}), & g^{(t,e)}_{\mathtt{B},i,r_i} &= \nabla_{\mathtt{B}}\mathcal{L}_i^{\text{lora}}(\mathtt{B}^{(t,e)}_{i,r_i}, \mathtt{A}^{(t,e)}_{i,r_i}) \\
\bar{g}^{(t)}_{\mathtt{A},i,r_i} &= \frac{1}{E}\sum_{e=0}^{E-1} \hat{g}^{(t,e)}_{\mathtt{A},i,r_i}, & \bar{g}^{(t)}_{\mathtt{B},i,r_i} &= \frac{1}{E}\sum_{e=0}^{E-1} \hat{g}^{(t,e)}_{\mathtt{B},i,r_i} \\
D^{(t)}_{\mathtt{A},i,r_i} &= \mathbb{E}_t\left[\bar{g}^{(t)}_{\mathtt{A},i,r_i}\right] - g^{(t)}_{\mathtt{A},i,r_i}, & D^{(t)}_{\mathtt{B},i,r_i} &= \mathbb{E}_t\left[\bar{g}^{(t)}_{\mathtt{B},i,r_i}\right] - g^{(t)}_{\mathtt{B},i,r_i}
\end{aligned}
$$

In what follows, we prove the necessary lemmas.

Lemma 2. *Under Assumption 1, the local empirical loss $\mathcal{L}_i^{full}(\Delta\mathtt{W})$, the global empirical loss $\mathcal{L}^{full}(\Delta\mathtt{W})$, and the stochastic gradient $\mathcal{L}_i^{full}(\Delta\mathtt{W}; \xi_i)$ evaluated on mini-batch ξ_i are μ-smooth with respect to $\Delta\mathtt{W}$, i.e.,*

$$
\begin{aligned}
\|\nabla_{\Delta\mathtt{W}}\mathcal{L}_i^{full}(\Delta\mathtt{W}) - \nabla_{\Delta\mathtt{W}}\mathcal{L}_i^{full}(\Delta\mathtt{W}')\|_F &\le \mu\|\Delta\mathtt{W} - \Delta\mathtt{W}'\|_F, \\
\|\nabla_{\Delta\mathtt{W}}\mathcal{L}^{full}(\Delta\mathtt{W}) - \nabla_{\Delta\mathtt{W}}\mathcal{L}^{full}(\Delta\mathtt{W}')\|_F &\le \mu\|\Delta\mathtt{W} - \Delta\mathtt{W}'\|_F, \quad \textit{and} \\
\mathbb{E}_{\xi_i}\left[\|\nabla_{\Delta\mathtt{W}}\mathcal{L}_i^{full}(\Delta\mathtt{W}; \xi_i) - \nabla_{\Delta\mathtt{W}}\mathcal{L}_i^{full}(\Delta\mathtt{W}'; \xi_i)\|_F\right] &\le \mu\|\Delta\mathtt{W} - \Delta\mathtt{W}'\|_F.
\end{aligned}
$$

Proof. Each loss in the statement is an average of sample losses. Averaging the inequality in Assumption 1 and using the triangle inequality concludes the proof. □

Lemma 3. *Under Assumption 2 and Assumption 4,* SEFORA *guarantees that, the local drift of the LoRA factors $\mathtt{B}$ and $\mathtt{A}$, and the update $\Delta\mathtt{W}$ at the local step $e \in \{0, \cdots, E-1\}$, and for any round $t \in \{0, \cdots, T-1\}$ and client $i \in [N]$ are bounded as:*

$$
\begin{aligned}
\mathbb{E}_t\left\|\mathtt{B}^{(t,e)}_{i,r_i} - \mathtt{B}^{(t,0)}_{i,r_i}\right\|_F^2 &\le (\eta^{(t)})^2e^2\chi^2M_{\mathbf{A}}^2, \\
\mathbb{E}_t\left\|\mathtt{A}^{(t,e)}_{i,r_i} - \mathtt{A}^{(t,0)}_{i,r_i}\right\|_F^2 &\le (\eta^{(t)})^2e^2\chi^2M_{\mathbf{B}}^2, \quad \textit{and} \\
\mathbb{E}_t\left\|\Delta\mathtt{W}^{(t,e)}_{i,r_i} - \Delta\mathtt{W}^{(t,0)}_{i,r_i}\right\|_F^2 &\le 2(\eta^{(t)})^2e^2\chi^2(M_{\mathbf{B}}^4 + M_{\mathbf{A}}^4).
\end{aligned}
$$

Proof. The local fine-tuning starts from $(\mathtt{B}_{i,r_i}^{(t,0)}, \mathtt{A}_{i,r_i}^{(t,0)})$ and updates

$$\mathtt{B}_{i,r_i}^{(t,e+1)} = \mathtt{B}_{i,r_i}^{(t,e)} - \eta^{(t)} \cdot \hat{g}_{\mathtt{B},i,r_i}^{(t,e)}$$
$$\mathtt{A}_{i,r_i}^{(t,e+1)} = \mathtt{A}_{i,r_i}^{(t,e)} - \eta^{(t)} \cdot \hat{g}_{\mathtt{A},i,r_i}^{(t,e)}$$

where we define $\hat{g}_{\mathtt{A},i,r_i}^{(t,e)} = \nabla_{\mathtt{A}} \mathcal{L}_i^{\text{full}}(\mathtt{B}_{i,r_i}^{(t,e)} \mathtt{A}_{i,r_i}^{(t,e)}; \xi_i^{(t,e)}) = (\mathtt{B}_{i,r_i}^{(t,e)})^\top \cdot \nabla_{\Delta \mathtt{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}_{i,r_i}^{(t,e)}; \xi_i^{(t,e)})$ and $\hat{g}_{\mathtt{B},i,r_i}^{(t,e)} = \nabla_{\mathtt{B}} \mathcal{L}_i^{\text{full}}(\mathtt{B}_{i,r_i}^{(t,e)} \mathtt{A}_{i,r_i}^{(t,e)}; \xi_i^{(t,e)}) = \nabla_{\Delta \mathtt{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}_{i,r_i}^{(t,e)}; \xi_i^{(t,e)}) \cdot (\mathtt{A}_{i,r_i}^{(t,e)})^\top$. Thus, for $e \in \{0, \cdots, E-1\}$,

$$\begin{aligned}
\mathbb{E}_t \left\| \mathtt{B}_{i,r_i}^{(t,e)} - \mathtt{B}_{i,r_i}^{(t,0)} \right\|_F^2 &= \mathbb{E}_t \left\| -\eta^{(t)} \sum_{j=0}^{e-1} \hat{g}_{\mathtt{B},i,r_i}^{(t,j)} \right\|_F^2 \\
&= \mathbb{E}_t \left[(\eta^{(t)})^2 \cdot \left\| \sum_{j=0}^{e-1} \hat{g}_{\mathtt{B},i,r_i}^{(t,j)} \right\|_F^2 \right] \\
&\overset{(a)}{\leq} (\eta^{(t)})^2 e \cdot \sum_{j=0}^{e-1} \mathbb{E}_t \left\| \hat{g}_{\mathtt{B},i,r_i}^{(t,j)} \right\|_F^2 \\
&\overset{(b)}{\leq} (\eta^{(t)})^2 e \cdot \sum_{j=0}^{e-1} \mathbb{E}_t \left\| \nabla_{\Delta \mathtt{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}_{i,r_i}^{(t,j)}; \xi_i^{(t,j)}) \right\|_F^2 \cdot \left\| \mathtt{A}_{i,r_i}^{(t,j)} \right\|_2^2 \\
&\overset{(c)}{\leq} (\eta^{(t)})^2 e^2 \chi^2 M_{\mathbf{A}}^2
\end{aligned}$$

where (a) holds due to the Jensen's inequality, (b) employs $\|XY\|_F^2 \leq \|X\|_2^2 \cdot \|Y\|_F^2$, and (c) utilizes Assumption 2 and Assumption 4. Similarly, $\mathbb{E}_t \left\| \mathtt{A}_{i,r_i}^{(t,e)} - \mathtt{A}_{i,r_i}^{(t,0)} \right\|_F^2 \leq (\eta^{(t)})^2 e^2 \chi^2 M_{\mathbf{B}}^2$. Then,

$$\begin{aligned}
\mathbb{E}_t \left\| \Delta \mathtt{W}_{i,r_i}^{(t,e)} - \Delta \mathtt{W}_{i,r_i}^{(t,0)} \right\|_F^2 &= \mathbb{E}_t \left\| \mathtt{B}_{i,r_i}^{(t,e)} \mathtt{A}_{i,r_i}^{(t,e)} - \mathtt{B}_{i,r_i}^{(t,0)} \mathtt{A}_{i,r_i}^{(t,0)} \right\|_F^2 \\
&= \mathbb{E}_t \left\| \mathtt{B}_{i,r_i}^{(t,e)} (\mathtt{A}_{i,r_i}^{(t,e)} - \mathtt{A}_{i,r_i}^{(t,0)}) + (\mathtt{B}_{i,r_i}^{(t,e)} - \mathtt{B}_{i,r_i}^{(t,0)}) \mathtt{A}_{i,r_i}^{(t,0)} \right\|_F^2 \\
&\leq 2\mathbb{E}_t \left\| \mathtt{B}_{i,r_i}^{(t,e)} \right\|_2^2 \left\| \mathtt{A}_{i,r_i}^{(t,e)} - \mathtt{A}_{i,r_i}^{(t,0)} \right\|_F^2 + 2\mathbb{E}_t \left\| \mathtt{B}_{i,r_i}^{(t,e)} - \mathtt{B}_{i,r_i}^{(t,0)} \right\|_F^2 \left\| \mathtt{A}_{i,r_i}^{(t,0)} \right\|_2^2 \\
&\leq 2M_{\mathbf{B}}^2 \mathbb{E}_t \left\| \mathtt{A}_{i,r_i}^{(t,e)} - \mathtt{A}_{i,r_i}^{(t,0)} \right\|_F^2 + 2M_{\mathbf{A}}^2 \mathbb{E}_t \left\| \mathtt{B}_{i,r_i}^{(t,e)} - \mathtt{B}_{i,r_i}^{(t,0)} \right\|_F^2 \\
&\leq 2(\eta^{(t)})^2 e^2 \chi^2 (M_{\mathbf{B}}^4 + M_{\mathbf{A}}^4)
\end{aligned}$$

□

Lemma 4. *For clients $i \in [N]$ and round $t \in \{0, \cdots, T-1\}$, under Assumption 1, Assumption 2 and Assumption 4, the following bounds hold:*

$$\sum_{i \in [N]} p_i \left\| D_{\mathtt{A},i,r_i}^{(t)} \right\|_F^2 \leq \frac{(E-1)(2E-1)}{3} \chi^2 (\eta^{(t)})^2 \big(\chi^2 M_{\mathbf{A}}^2 + 2\mu^2 M_{\mathbf{B}}^2 (M_{\mathbf{B}}^4 + M_{\mathbf{A}}^4) \big),$$

and

$$\sum_{i \in [N]} p_i \left\| D_{\mathtt{B},i,r_i}^{(t)} \right\|_F^2 \leq \frac{(E-1)(2E-1)}{3} \chi^2 (\eta^{(t)})^2 \big(\chi^2 M_{\mathbf{B}}^2 + 2\mu^2 M_{\mathbf{A}}^2 (M_{\mathbf{B}}^4 + M_{\mathbf{A}}^4) \big).$$

Proof. We prove the bound on $\mathtt{A}$; the $\mathtt{B}$ bound follows by the symmetric argument. The Frobenius norm of the drifts $\|D_{\mathtt{A},i,r_i}^{(t)}\|_F^2$ is bounded as:

$$\left\| D_{\mathtt{A},i,r_i}^{(t)} \right\|_F^2 = \left\| \mathbb{E}_t \left[\frac{1}{E} \sum_{e=0}^{E-1} \hat{g}_{\mathtt{A},i,r_i}^{(t,e)} \right] - g_{\mathtt{A},i,r_i}^{(t)} \right\|_F^2$$

$$
\begin{aligned}
&= \left\| \frac{1}{E} \sum_{e=0}^{E-1} \mathbb{E}_t \left[\hat{g}^{(t,e)}_{\mathtt{A},i,r_i} - g^{(t)}_{\mathtt{A},i,r_i} \right] \right\|_F^2 \\
&\overset{(a)}{=} \left\| \frac{1}{E} \sum_{e=0}^{E-1} \mathbb{E}_t \left[g^{(t,e)}_{\mathtt{A},i,r_i} - g^{(t)}_{\mathtt{A},i,r_i} \right] \right\|_F^2 \\
&\overset{(b)}{\leq} \frac{1}{E} \sum_{e=0}^{E-1} \mathbb{E}_t \left\| g^{(t,e)}_{\mathtt{A},i,r_i} - g^{(t)}_{\mathtt{A},i,r_i} \right\|_F^2 \\
&= \frac{1}{E} \sum_{e=0}^{E-1} \mathbb{E}_t \left\| (\mathtt{B}^{(t,e)}_{i,r_i})^\top \cdot \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}_i (\Delta \mathtt{W}^{(t,e)}_{i,r_i}) - (\mathtt{B}^{(t,0)}_{i,r_i})^\top \cdot \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}_i (\Delta \mathtt{W}^{(t,0)}_{i,r_i}) \right\|_F^2 \\
&= \frac{1}{E} \sum_{e=0}^{E-1} \mathbb{E}_t \Bigg\| \left((\mathtt{B}^{(t,e)}_{i,r_i})^\top - (\mathtt{B}^{(t,0)}_{i,r_i})^\top \right) \cdot \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}_i (\Delta \mathtt{W}^{(t,e)}_{i,r_i}) \\
&\qquad\qquad + (\mathtt{B}^{(t,0)}_{i,r_i})^\top \cdot \left(\nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}_i (\Delta \mathtt{W}^{(t,e)}_{i,r_i}) - \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}_i (\Delta \mathtt{W}^{(t,0)}_{i,r_i}) \right) \Bigg\|_F^2 \\
&\overset{(c)}{\leq} 2 \frac{1}{E} \sum_{e=0}^{E-1} \mathbb{E}_t \left\| (\mathtt{B}^{(t,e)}_{i,r_i})^\top - (\mathtt{B}^{(t,0)}_{i,r_i})^\top \right\|_F^2 \left\| \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}_i (\Delta \mathtt{W}^{(t,e)}_{i,r_i}) \right\|_F^2 \\
&\quad + 2 \frac{1}{E} \sum_{e=0}^{E-1} \mathbb{E}_t \left\| (\mathtt{B}^{(t,0)}_{i,r_i})^\top \right\|_2^2 \left\| \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}_i (\Delta \mathtt{W}^{(t,e)}_{i,r_i}) - \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}_i (\Delta \mathtt{W}^{(t,0)}_{i,r_i}) \right\|_F^2 \\
&\overset{(d)}{\leq} 2\chi^2 \frac{1}{E} \sum_{e=0}^{E-1} \mathbb{E}_t \left\| (\mathtt{B}^{(t,e)}_{i,r_i})^\top - (\mathtt{B}^{(t,0)}_{i,r_i})^\top \right\|_F^2 + 2 \cdot \mu^2 M_{\mathbf{B}}^2 \frac{1}{E} \sum_{e=0}^{E-1} \mathbb{E}_t \left\| \Delta \mathtt{W}^{(t,e)}_{i,r_i} - \Delta \mathtt{W}^{(t,0)}_{i,r_i} \right\|_F^2 \\
&\overset{(e)}{\leq} 2\chi^2 \frac{1}{E} \sum_{e=0}^{E-1} (\eta^{(t)})^2 e^2 \chi^2 M_{\mathbf{A}}^2 + 2 \cdot \mu^2 M_{\mathbf{B}}^2 \frac{1}{E} \sum_{e=0}^{E-1} 2 (\eta^{(t)})^2 e^2 \chi^2 (M_{\mathbf{B}}^4 + M_{\mathbf{A}}^4) \\
&= 2\chi^2 (\eta^{(t)})^2 \left(\chi^2 M_{\mathbf{A}}^2 + 2\mu^2 M_{\mathbf{B}}^2 (M_{\mathbf{B}}^4 + M_{\mathbf{A}}^4) \right) \cdot \frac{1}{E} \sum_{e=0}^{E-1} e^2 \\
&= \frac{(E-1)(2E-1)}{3} \chi^2 (\eta^{(t)})^2 \left(\chi^2 M_{\mathbf{A}}^2 + 2\mu^2 M_{\mathbf{B}}^2 (M_{\mathbf{B}}^4 + M_{\mathbf{A}}^4) \right)
\end{aligned}
$$

where (a) holds since $\mathbb{E}_t \left[g^{(t,e)}_{\mathtt{A},i,r_i} \right] = \mathbb{E}_t \left[\hat{g}^{(t,e)}_{\mathtt{A},i,r_i} \right]$, which is obtained by conditioning on the history before sampling of the mini-batch at local step e and applying the tower property, (b) holds by applying Jensen's inequality, (c) holds since $\|XY\|_F^2 \leq \|X\|_2^2 \cdot \|Y\|_F^2 \leq \|X\|_F^2 \cdot \|Y\|_F^2$, (d) holds since $\left\| \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}_i (\Delta \mathtt{W}^{(t,e)}_i) \right\|_F^2 = \left\| \mathbb{E}_{\xi_i} \left[\nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}_i (\Delta \mathtt{W}^{(t,e)}_i; \xi_i) \right] \right\|_F^2 \leq \mathbb{E}_{\xi_i} \left\| \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}_i (\Delta \mathtt{W}^{(t,e)}_i; \xi_i) \right\|_F^2 \leq \chi^2$, Assumption 1 and Assumption 4, and (e) holds due to Lemma 3 and $\sum_{i=0}^{E-1} e^2 = \frac{(E-1)E(2E-1)}{6}$. Similarly,

$$
\|D^{(t)}_{\mathtt{B},i,r_i}\|_F^2 \leq \frac{(E-1)(2E-1)}{3} \chi^2 (\eta^{(t)})^2 \left(\chi^2 M_{\mathbf{B}}^2 + 2\mu^2 M_{\mathbf{A}}^2 (M_{\mathbf{B}}^4 + M_{\mathbf{A}}^4) \right).
$$

Because $\sum_i p_i = 1$, the same bounds hold for $\sum_i p_i \|D^{(t)}_{\mathtt{A},i,r_i}\|_F^2$ and $\sum_i p_i \|D^{(t)}_{\mathtt{B},i,r_i}\|_F^2$. □

Lemma 5. *For each client $i \in [N]$ and round $t \in \{0, \cdots, T-1\}$, under Assumption 2 and Assumption 4, the following bounds hold:* $\mathbb{E}_t \left\| \bar{g}^{(t)}_{\mathtt{B},i,r_i} \bar{g}^{(t)}_{\mathtt{A},i,r_i} \right\|_F^2 \leq M_{\mathbf{A}}^2 M_{\mathbf{B}}^2 \kappa^4$.

Proof. By definition, $\bar{g}^{(t)}_{\mathtt{B},i,r_i} \bar{g}^{(t)}_{\mathtt{A},i,r_i} = \frac{1}{E^2} \sum_{e=0}^{E-1} \hat{g}^{(t,e)}_{\mathtt{B},i,r_i} \sum_{e=0}^{E-1} \hat{g}^{(t,e)}_{\mathtt{A},i,r_i}$, thus

$$
\mathbb{E}_t \left\| \bar{g}^{(t)}_{\mathtt{B},i,r_i} \bar{g}^{(t)}_{\mathtt{A},i,r_i} \right\|_F^2 = \mathbb{E}_t \left[\left\| \frac{1}{E^2} \sum_{e=0}^{E-1} \hat{g}^{(t,e)}_{\mathtt{B},i,r_i} \sum_{e=0}^{E-1} \hat{g}^{(t,e)}_{\mathtt{A},i,r_i} \right\|_F^2 \right]
$$

$$
\begin{aligned}
&\overset{(a)}{\leq} \frac{1}{E^4} E^2 \cdot \sum_{e_1,e_2} \mathbb{E}_t \left[\left\| \hat{g}^{(t,e_1)}_{\mathtt{B},i,r_i} \cdot \hat{g}^{(t,e_2)}_{\mathtt{A},i,r_i} \right\|_F^2 \right] \\
&= \frac{1}{E^2} \cdot \sum_{e_1,e_2} \mathbb{E}_t \left[\left\| \nabla_{\Delta \mathtt{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}^{(t,e_1)}_{i,r_i}; \xi_i^{(t,e_1)}) \cdot (\mathtt{A}^{(t,e_1)}_{i,r_i})^\top \cdot (\mathtt{B}^{(t,e_2)}_{i,r_i})^\top \cdot \nabla_{\Delta \mathtt{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}^{(t,e_2)}_{i,r_i}; \xi_i^{(t,e_2)}) \right\| \right] \\
&= \frac{1}{E^2} \sum_{e_1,e_2} \mathbb{E}_t \Bigg[\sum_{l=1}^{L} \Big\| \nabla_{\Delta \mathbf{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathbf{W}^{(t,e_1)}_{i,l,r_{i,l}}; \xi_i^{(t,e_1)}) (\mathbf{A}^{(t,e_1)}_{i,l,r_{i,l}})^\top \cdot \\
&\qquad\qquad (\mathbf{B}^{(t,e_2)}_{i,l,r_{i,l}})^\top \nabla_{\Delta \mathbf{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathbf{W}^{(t,e_2)}_{i,l,r_{i,l}}; \xi_i^{(t,e_2)}) \Big\|_F^2 \Bigg] \\
&\overset{(b)}{\leq} \frac{1}{E^2} \sum_{e_1,e_2} \mathbb{E}_t \Bigg[\sum_{l=1}^{L} \left\| \nabla_{\Delta \mathbf{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathbf{W}^{(t,e_1)}_{i,l,r_{i,l}}; \xi_i^{(t,e_1)}) \right\|_F^2 \cdot \left\| \mathbf{A}^{(t,e_1)}_{i,l,r_{i,l}} \right\|_2^2 \cdot \\
&\qquad\qquad \left\| \mathbf{B}^{(t,e_2)}_{i,l,r_{i,l}} \right\|_2^2 \cdot \left\| \nabla_{\Delta \mathbf{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathbf{W}^{(t,e_2)}_{i,l,r_{i,l}}; \xi_i^{(t,e_2)}) \right\|_F^2 \Bigg] \\
&\overset{(c)}{\leq} \frac{M_{\mathbf{A}}^2 M_{\mathbf{B}}^2}{E^2} \sum_{e_1,e_2} \mathbb{E}_t \left[\sum_{l=1}^{L} \left\| \nabla_{\Delta \mathbf{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathbf{W}^{(t,e_1)}_{i,l,r_{i,l}}; \xi_i^{(t,e_1)}) \right\|_F^2 \cdot \left\| \nabla_{\Delta \mathbf{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathbf{W}^{(t,e_2)}_{i,l,r_{i,l}}; \xi_i^{(t,e_2)}) \right\|_F^2 \right] \\
&\overset{(d)}{\leq} \frac{M_{\mathbf{A}}^2 M_{\mathbf{B}}^2}{E^2} \sum_{e_1,e_2} \mathbb{E}_t \left[\left\| \nabla_{\Delta \mathtt{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}^{(t,e_1)}_{i,r_i}; \xi_i^{(t,e_1)}) \right\|_F^2 \left\| \nabla_{\Delta \mathtt{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}^{(t,e_2)}_{i,r_i}; \xi_i^{(t,e_2)}) \right\|_F^2 \right] \\
&\overset{(e)}{\leq} \frac{M_{\mathbf{A}}^2 M_{\mathbf{B}}^2}{E^2} \sum_{e_1,e_2} \sqrt{\mathbb{E}_t \left[\left\| \nabla_{\Delta \mathtt{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}^{(t,e_1)}_{i,r_i}; \xi_i^{(t,e_1)}) \right\|_F^4 \right] \mathbb{E}_t \left[\left\| \nabla_{\Delta \mathtt{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}^{(t,e_2)}_{i,r_i}; \xi_i^{(t,e_2)}) \right\|_F^4 \right]} \\
&\overset{(f)}{\leq} \frac{M_{\mathbf{A}}^2 M_{\mathbf{B}}^2}{E^2} \sum_{e_1,e_2} \sqrt{\kappa^4 \cdot \kappa^4} \\
&\leq M_{\mathbf{A}}^2 M_{\mathbf{B}}^2 \kappa^4,
\end{aligned}
$$

where in (a) holds due to Jensen's inequality, in (b) we employ $\|XY\|_F^2 \leq \|X\|_2^2 \cdot \|Y\|_F^2 \leq \|X\|_F^2 \cdot \|Y\|_F^2$, (c) holds due to Assumption 4, (d) holds because $\sum x_l^2 y_l^2 \leq \sum x_l^2 \sum y_l^2$ for $x_l \geq 0$, (e) holds due to Cauchy-Schwarz inequality, and (f) uses Assumption 2. □

Lemma 6. *For each client $i \in [N]$ and round $t \in \{0, \cdots, T-1\}$, under Assumption 2 and Assumption 4,* $\mathbb{E}_t \left\| \bar{g}^{(t)}_{\mathtt{A},i,r_i} \right\|_F^2 \leq M_{\mathbf{B}}^2 \chi^2$ *and* $\mathbb{E}_t \left\| \bar{g}^{(t)}_{\mathtt{B},i,r_i} \right\|_F^2 \leq M_{\mathbf{A}}^2 \chi^2$.

Proof. We prove $\mathbb{E}_t \|\bar{g}^{(t)}_{\mathtt{A},i,r_i}\|_F^2 \leq M_{\mathbf{B}}^2 \chi^2$ below. The bound on $\bar{g}^{(t)}_{\mathtt{B},i}$ is symmetric.

$$
\begin{aligned}
\mathbb{E}_t \left\| \bar{g}^{(t)}_{\mathtt{A},i,r_i} \right\|_F^2 &= \frac{1}{E^2} \mathbb{E}_t \left\| \sum_{e=0}^{E-1} \hat{g}^{(t,e)}_{\mathtt{A},i,r_i} \right\|_F^2 \\
&\overset{(a)}{\leq} \frac{1}{E^2} E \sum_{e=0}^{E-1} \mathbb{E}_t \left\| \hat{g}^{(t,e)}_{\mathtt{A},i,r_i} \right\|_F^2 \\
&= \frac{1}{E} \sum_{e=0}^{E-1} \mathbb{E}_t \left\| (\mathtt{B}^{(t,e)}_{i,r_i})^\top \cdot \nabla_{\Delta \mathtt{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}^{(t,e)}_{i,r_i}; \xi_i^{(t,e)}) \right\|_F^2 \\
&\overset{(b)}{\leq} \frac{1}{E} \sum_{e=0}^{E-1} \|\mathtt{B}^{(t,e)}_{i,r_i}\|_2^2 \mathbb{E}_t \left\| \nabla_{\Delta \mathtt{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}^{(t,e)}_{i,r_i}; \xi_i^{(t,e)}) \right\|_F^2 \\
&\overset{(c)}{\leq} M_{\mathbf{B}}^2 \chi^2,
\end{aligned}
$$

where (a) is Jensen's inequality, (b) uses $\|XY\|_F^2 \leq \|X\|_2^2 \|Y\|_F^2$, and (c) applies Assumption 4 and Assumption 2. The bound on $\bar{g}^{(t)}_{\mathtt{B},i,r_i}$ is identical with the roles of $\mathtt{A}$ and $\mathtt{B}$ swapped. □

E.1 Proof of Theorem 1

Proof. At round $t-1$ ($t \geq 1$[3]), suppose the unsketching algorithm (Algorithm 2) returns $\widehat{\mathbf{B}}_l^{(t)}$ and $\widehat{\mathbf{A}}_l^{(t)}$ for all layer $l \in [L]$. The corresponding global update is $\Delta\mathbf{W}_l^{(t)} = \widehat{\mathbf{B}}_l^{(t)}\widehat{\mathbf{A}}_l^{(t)}$, with global rank r_l. Let $\widehat{\mathbf{A}}_l^{(t)} = \mathbf{U}_l^{(t)}\mathbf{\Sigma}_l^{(t)}(\mathbf{V}_l^{(t)})^\top$ be its SVD. Since both $\widehat{\mathbf{B}}_l^{(t)}$ and $\mathbf{U}_l^{(t)}$ have orthonormal columns, $\Delta\mathbf{W}_l^{(t)} = \widehat{\mathbf{B}}_l^{(t)}\mathbf{U}_l^{(t)}\mathbf{\Sigma}_l^{(t)}(\mathbf{V}_l^{(t)})^\top$ is an SVD of $\Delta\mathbf{W}_l^{(t)}$. We consequently define the global factors as $\mathbf{B}_l^{(t)} = \widehat{\mathbf{B}}_l^{(t)}\mathbf{U}_l^{(t)}(\mathbf{\Sigma}_l^{(t)})^{\frac{1}{2}}$ and $\mathbf{A}_l^{(t)} = (\mathbf{\Sigma}_l^{(t)})^{\frac{1}{2}}(\mathbf{V}_l^{(t)})^\top$, which is computed and broadcast by the federator. Since client i's local rank $r_{i,l} \leq r_l$, according to line 4 of Algorithm 1, computing the initial local LoRA factors $\mathbf{B}_{i,l,r_{i,l}}^{(t,0)} \in \mathbb{R}^{d\times r_{i,l}}$ and $\mathbf{A}_{i,l,r_{i,l}}^{(t,0)} \in \mathbb{R}^{r_{i,l}\times s}$ is simply to perform truncation through forming the submatrices of the SVD decomposition of matrix $\widehat{\mathbf{A}}_l^{(t)} = \mathbf{U}_l^{(t)}\mathbf{\Sigma}_l^{(t)}(\mathbf{V}_l^{(t)})^\top$, i.e.,

$$\begin{aligned}\mathbf{B}_{i,l,r_{i,l}}^{(t,0)} &\leftarrow \widehat{\mathbf{B}}_l^{(t)}(\mathbf{U}_l^{(t)})_{:,1:r_{i,l}}(\mathbf{\Sigma}_l^{(t)})^{1/2}_{1:r_{i,l},1:r_{i,l}},\\ \mathbf{A}_{i,l,r_{i,l}}^{(t,0)} &\leftarrow (\mathbf{\Sigma}_l^{(t)})^{1/2}_{1:r_{i,l},1:r_{i,l}}(\mathbf{V}_l^{(t)})^\top_{1:r_{i,l},:}\,.\end{aligned}$$

Whenever a quantity is indexed by both i and l, its local rank is $r_{i,l}$; we suppress the rank subscript when no ambiguity arises. Thus, we have

$$\mathbf{B}_{i,l}^{(t,0)} = \mathbf{B}_l^{(t)}\mathbf{\Gamma}_{i,l}\,, \quad \mathbf{A}_{i,l}^{(t,0)} = (\mathbf{\Gamma}_{i,l})^\top\mathbf{A}_l^{(t)}, \tag{1}$$

and

$$\Delta\mathbf{W}_{i,l}^{(t,0)} = \mathbf{B}_{i,l}^{(t,0)}\mathbf{A}_{i,l}^{(t,0)} = \mathbf{B}_l^{(t)}\mathbf{\Gamma}_{i,l}(\mathbf{\Gamma}_{i,l})^\top\mathbf{A}_l^{(t)} = \mathfrak{b}(\Delta\mathbf{W}_l^{(t)}).$$

Notice that, $\mathbf{\Gamma}_{i,l}^\top\mathbf{\Gamma}_{i,l} = \mathbf{I}_{r_{i,l}}$ and $\|\mathbf{\Gamma}_{i,l}\mathbf{\Gamma}_{i,l}^\top\|_2 = 1$. For $t=0$, Equation (1) also holds since each client obtains its initial local LoRA factors by truncating $\mathbf{B}_l^{(0)}$ and $\mathbf{A}_l^{(0)}$. Thus, the subsequent proof applies to every $t \in \{0,\cdots,T-1\}$.

Let the global model produced by SeFoRA for round $t+1$ be $\Delta\mathtt{W}_{\text{SeFoRA}}^{(t+1)}$. Since the full objective is μ-smooth with respect to $\Delta\mathtt{W}$, we have

$$\begin{aligned}&\mathbb{E}_t\left[\mathcal{L}^{\text{full}}(\Delta\mathtt{W}_{\text{SeFoRA}}^{(t+1)})\right] \qquad (2)\\ \leq&\mathbb{E}_t\left[\mathcal{L}^{\text{full}}(\Delta\mathtt{W}^{(t)})\right] + \underbrace{\mathbb{E}_t\left[\left\langle\nabla_{\Delta\mathtt{W}}\mathcal{L}^{\text{full}}(\Delta\mathtt{W}^{(t)}), \Delta\mathtt{W}_{\text{SeFoRA}}^{(t+1)} - \Delta\mathtt{W}^{(t)}\right\rangle\right]}_{\text{Term 1}} + \underbrace{\frac{\mu}{2}\mathbb{E}_t\left[\|\Delta\mathtt{W}_{\text{SeFoRA}}^{(t+1)} - \Delta\mathtt{W}^{(t)}\|_F^2\right]}_{\text{Term 2}},\end{aligned}$$

Suppose the global LoRA factors at round t are $\mathtt{A}^{(t)}$ and $\mathtt{B}^{(t)}$, under the rank-heterogeneous scenarios, each client i first truncates the global LoRA factors to local rank r_i, i.e., $[\![\mathtt{A}^{(t)}]\!]_{r_i}$ and $[\![\mathtt{B}^{(t)}]\!]_{r_i}$, before proceeding with local training. The local fine-tuning starts from $(\mathtt{B}_{i,r_i}^{(t,0)}, \mathtt{A}_{i,r_i}^{(t,0)}) = ([\![\mathtt{B}^{(t)}]\!]_{r_i}, [\![\mathtt{A}^{(t)}]\!]_{r_i})$ and updates

$$\begin{aligned}\mathtt{B}_{i,r_i}^{(t,e+1)} &= \mathtt{B}_{i,r_i}^{(t,e)} - \eta^{(t)}\hat{g}_{\mathtt{B},i,r_i}^{(t,e)}, \quad \text{and}\\ \mathtt{A}_{i,r_i}^{(t,e+1)} &= \mathtt{A}_{i,r_i}^{(t,e)} - \eta^{(t)}\hat{g}_{\mathtt{A},i,r_i}^{(t,e)}.\end{aligned}$$

Thus,

$$\begin{aligned}\mathtt{B}_{i,r_i}^{(t,E)} &= \mathtt{B}_{i,r_i}^{(t,0)} - \eta^{(t)}\sum_{e=0}^{E-1}\hat{g}_{\mathtt{B},i,r_i}^{(t,e)} = \mathtt{B}_{i,r_i}^{(t,0)} - \eta^{(t)}E\cdot\bar{g}_{\mathtt{B},i,r_i}^{(t)}, \quad \text{and}\\ \mathtt{A}_{i,r_i}^{(t,E)} &= \mathtt{A}_{i,r_i}^{(t,0)} - \eta^{(t)}\sum_{e=0}^{E-1}\hat{g}_{\mathtt{A},i,r_i}^{(t,e)} = \mathtt{A}_{i,r_i}^{(t,0)} - \eta^{(t)}E\cdot\bar{g}_{\mathtt{A},i,r_i}^{(t)}.\end{aligned}$$

Define $\mathtt{V}^{(t+1)} = \Delta\mathtt{W}^{(t)} + \sum_{i\in[N]}p_i(\Delta\mathtt{W}_{i,r_i}^{(t,E)} - \Delta\mathtt{W}_{i,r_i}^{(t,0)})$. In round t and for each layer $l \in [L]$, the federator aggregates $\mathbf{Y}_l^{(t+1)} = \mathbf{Y}_{\text{glo},l}^{(t)} + \sum_{i\in[N]}p_i\mathbf{Y}_{i,l}^{(t)}$ and $\mathbf{Z}_l^{(t+1)} = \mathbf{Z}_{\text{glo},l}^{(t)} + \sum_{i\in[N]}p_i\mathbf{Z}_{i,l}^{(t)}$,

[3]Round 0 uses the original initialization, where $\mathbf{A}_l^{(0)}$ with random Gaussian entries and $\mathbf{B}_l^{(0)}$ as the zero matrix.

which are two sketches of $\mathbf{V}_l^{(t+1)}$. Therefore, after unsketching, the updated model for round $t+1$ is $\Delta \mathtt{W}_{\textsc{SeFoRA}}^{(t+1)} = \mathtt{V}^{(t+1)} + \mathcal{E}_{\text{sketch}}^{(t+1)}$, where $\mathcal{E}_{\text{sketch}}^{(t+1)}$ is the sketching error with $\mathtt{V}^{(t+1)}$ being the target tuple. We have

$$\begin{aligned}\Delta \mathtt{W}_{\textsc{SeFoRA}}^{(t+1)} &= \mathtt{V}^{(t+1)} + \mathcal{E}_{\text{sketch}}^{(t+1)} \\ &= \Delta \mathtt{W}^{(t)} + \sum_{i\in[N]} p_i (\Delta \mathtt{W}_{i,r_i}^{(t,E)} - \Delta \mathtt{W}_{i,r_i}^{(t,0)}) + \mathcal{E}_{\text{sketch}}^{(t+1)}\end{aligned}$$

Define $\delta_{i,r_i}^{(t)} \triangleq \Delta \mathtt{W}_{i,r_i}^{(t,E)} - \Delta \mathtt{W}_{i,r_i}^{(t,0)}$, we have

$$\Delta \mathtt{W}_{\textsc{SeFoRA}}^{(t+1)} - \Delta \mathtt{W}^{(t)} = \sum_{i\in[N]} p_i \delta_{i,r_i}^{(t)} + \mathcal{E}_{\text{sketch}}^{(t+1)}, \tag{3}$$

and

$$\begin{aligned}\delta_{i,r_i}^{(t)} &= \mathtt{B}_{i,r_i}^{(t,E)} \mathtt{A}_{i,r_i}^{(t,E)} - \mathtt{B}_{i,r_i}^{(t,0)} \mathtt{A}_{i,r_i}^{(t,0)} \\ &= (\mathtt{B}_{i,r_i}^{(t,0)} - \eta^{(t)} E \cdot \bar{g}_{\mathtt{B},i,r_i}^{(t)})(\mathtt{A}_{i,r_i}^{(t,0)} - \eta^{(t)} E \cdot \bar{g}_{\mathtt{A},i,r_i}^{(t)}) - \mathtt{B}_{i,r_i}^{(t,0)} \mathtt{A}_{i,r_i}^{(t,0)} \\ &= -\eta^{(t)} E \cdot \mathtt{B}_{i,r_i}^{(t,0)} \bar{g}_{\mathtt{A},i,r_i}^{(t)} - \eta^{(t)} E \cdot \bar{g}_{\mathtt{B},i,r_i}^{(t)} \mathtt{A}_{i,r_i}^{(t,0)} + (\eta^{(t)})^2 E^2 \cdot \bar{g}_{\mathtt{B},i,r_i}^{(t)} \bar{g}_{\mathtt{A},i,r_i}^{(t)} \\ &= -\eta^{(t)} E \cdot \mathtt{B}^{(t)} \Gamma_i \cdot \bar{g}_{\mathtt{A},i,r_i}^{(t)} - \eta^{(t)} E \cdot \bar{g}_{\mathtt{B},i,r_i}^{(t)} \cdot (\Gamma_i)^\top \mathtt{A}^{(t)} + (\eta^{(t)})^2 E^2 \cdot \bar{g}_{\mathtt{B},i,r_i}^{(t)} \bar{g}_{\mathtt{A},i,r_i}^{(t)}\end{aligned}$$

For Term 1:

$$\begin{aligned}\mathbb{E}_t\left[\left\langle \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)}), \Delta \mathtt{W}_{\textsc{SeFoRA}}^{(t+1)} - \Delta \mathtt{W}^{(t)} \right\rangle\right] &= \underbrace{\mathbb{E}_t\left[\left\langle \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)}), \sum_{i\in[N]} p_i \delta_{i,r_i}^{(t)} \right\rangle\right]}_{\text{Part A}} \\ &\quad + \underbrace{\mathbb{E}_t\left[\left\langle \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)}), \mathcal{E}_{\text{sketch}}^{(t+1)} \right\rangle\right]}_{\text{Part B}}\end{aligned}$$

We start by bounding Part A:

$$\begin{aligned}&\mathbb{E}_t\left[\left\langle \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)}), \sum_{i\in[N]} p_i \delta_{i,r_i}^{(t)} \right\rangle\right] \\ &= -\eta^{(t)} E \cdot \sum_{i\in[N]} p_i \mathbb{E}_t\left[\left\langle \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)}), \mathtt{B}^{(t)} \Gamma_i \cdot \bar{g}_{\mathtt{A},i,r_i}^{(t)} \right\rangle\right] \\ &\quad - \eta^{(t)} E \cdot \sum_{i\in[N]} p_i \mathbb{E}_t\left[\left\langle \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)}), \bar{g}_{\mathtt{B},i,r_i}^{(t)} \cdot \Gamma_i^\top \mathtt{A}^{(t)} \right\rangle\right] \\ &\quad + (\eta^{(t)})^2 E^2 \cdot \sum_{i\in[N]} p_i \mathbb{E}_t\left[\left\langle \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)}), \bar{g}_{\mathtt{B},i,r_i}^{(t)} \bar{g}_{\mathtt{A},i,r_i}^{(t)} \right\rangle\right] \\ &= -\eta^{(t)} E \cdot \sum_{i\in[N]} p_i \left\langle \Gamma_i^\top g_{\mathtt{A}}^{(t)}, D_{\mathtt{A},i,r_i}^{(t)} + g_{\mathtt{A},i,r_i}^{(t)} \right\rangle - \eta^{(t)} E \cdot \sum_{i\in[N]} p_i \left\langle g_{\mathtt{B}}^{(t)} \Gamma_i, D_{\mathtt{B},i,r_i}^{(t)} + g_{\mathtt{B},i,r_i}^{(t)} \right\rangle \\ &\quad + (\eta^{(t)})^2 E^2 \cdot \left\langle \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)}), \sum_{i\in[N]} p_i \mathbb{E}_t\left[\bar{g}_{\mathtt{B},i,r_i}^{(t)} \bar{g}_{\mathtt{A},i,r_i}^{(t)}\right] \right\rangle \\ &= -\eta^{(t)} E \cdot \sum_{i\in[N]} p_i \left\langle \Gamma_i^\top g_{\mathtt{A}}^{(t)}, g_{\mathtt{A},i,r_i}^{(t)} \right\rangle - \eta^{(t)} E \cdot \sum_{i\in[N]} p_i \left\langle g_{\mathtt{B}}^{(t)} \Gamma_i, g_{\mathtt{B},i,r_i}^{(t)} \right\rangle \\ &\quad - \eta^{(t)} E \cdot \sum_{i\in[N]} p_i \left\langle \Gamma_i^\top g_{\mathtt{A}}^{(t)}, D_{\mathtt{A},i,r_i}^{(t)} \right\rangle - \eta^{(t)} E \cdot \sum_{i\in[N]} p_i \left\langle g_{\mathtt{B}}^{(t)} \Gamma_i, D_{\mathtt{B},i,r_i}^{(t)} \right\rangle\end{aligned}$$

$$+ (\eta^{(t)})^2 E^2 \cdot \left\langle \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)}), \sum_{i\in[N]} p_i \mathbb{E}_t \left[\bar{g}^{(t)}_{\mathtt{B},i,r_i} \bar{g}^{(t)}_{\mathtt{A},i,r_i} \right] \right\rangle$$

where (a) holds since the definition of the tuple inner product and the cyclic property of the trace, i.e.,

$$\left\langle \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)}), \mathtt{B}^{(t)} \Gamma_i \cdot \bar{g}^{(t)}_{\mathtt{A},i,r_i} \right\rangle = \left\langle \Gamma_i^\top g^{(t)}_{\mathtt{A}}, \bar{g}^{(t)}_{\mathtt{A},i,r_i} \right\rangle,$$

$$\left\langle \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)}), \bar{g}^{(t)}_{\mathtt{B},i,r_i} \cdot \Gamma_i^\top \mathtt{A}^{(t)} \right\rangle = \left\langle g^{(t)}_{\mathtt{B}} \Gamma_i, \bar{g}^{(t)}_{\mathtt{B},i,r_i} \right\rangle,$$

and the definition of $D^{(t)}_{\mathtt{A}}$ and $D^{(t)}_{\mathtt{B}}$. We further write

$$g^{(t)}_{\mathtt{A},i,r_i} = (\mathtt{B}^{(t,0)}_{i,r_i})^\top \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}_i(\Delta \mathtt{W}^{(t,0)}_{i,r_i}) = (\mathtt{B}^{(t)} \Gamma_i)^\top \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}_i(\Delta \mathtt{W}^{(t,0)}_{i,r_i}),$$

$$g^{(t)}_{\mathtt{B},i,r_i} = \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}_i(\Delta \mathtt{W}^{(t,0)}_{i,r_i}) (\mathtt{A}^{(t,0)}_{i,r_i})^\top = \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}_i(\Delta \mathtt{W}^{(t,0)}_{i,r_i}) ((\Gamma_i)^\top \mathtt{A}^{(t)})^\top,$$

and

$$\nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}_i(\Delta \mathtt{W}^{(t,0)}_{i,r_i}) = \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)}) + (\nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}_i(\Delta \mathtt{W}^{(t,0)}_{i,r_i}) - \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)})),$$

thus, we have

$$\begin{aligned}
\left\langle \Gamma_i^\top g^{(t)}_{\mathtt{A}}, g^{(t)}_{\mathtt{A},i,r_i} \right\rangle &= \left\langle \Gamma_i^\top g^{(t)}_{\mathtt{A}}, (\mathtt{B}^{(t)} \Gamma_i)^\top \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}_i(\Delta \mathtt{W}^{(t,0)}_{i,r_i}) \right\rangle \\
&= \left\langle \Gamma_i^\top g^{(t)}_{\mathtt{A}}, \Gamma_i^\top (\mathtt{B}^{(t)})^\top \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)}) \right\rangle \\
&\quad + \left\langle \Gamma_i^\top g^{(t)}_{\mathtt{A}}, (\mathtt{B}^{(t)} \Gamma_i)^\top (\nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}_i(\Delta \mathtt{W}^{(t,0)}_{i,r_i}) - \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)})) \right\rangle \\
&= \left\| \Gamma_i^\top g^{(t)}_{\mathtt{A}} \right\|_F^2 + \left\langle \Gamma_i^\top g^{(t)}_{\mathtt{A}}, (\mathtt{B}^{(t)} \Gamma_i)^\top (\nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}_i(\Delta \mathtt{W}^{(t,0)}_{i,r_i}) - \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)})) \right\rangle \\
&= \left\| \Gamma_i^\top g^{(t)}_{\mathtt{A}} \right\|_F^2 + \left\langle \Gamma_i^\top g^{(t)}_{\mathtt{A}}, (\mathtt{B}^{(t)} \Gamma_i)^\top (\nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}_i(\Delta \mathtt{W}^{(t,0)}_{i,r_i}) - \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}_i(\Delta \mathtt{W}^{(t)})) \right\rangle \\
&\quad + \left\langle \Gamma_i^\top g^{(t)}_{\mathtt{A}}, (\mathtt{B}^{(t)} \Gamma_i)^\top (\nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}_i(\Delta \mathtt{W}^{(t)}) - \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)})) \right\rangle \\
\left\langle g^{(t)}_{\mathtt{B}} \Gamma_i, g^{(t)}_{\mathtt{B},i,r_i} \right\rangle &= \left\| g^{(t)}_{\mathtt{B}} \Gamma_i \right\|_F^2 + \left\langle g^{(t)}_{\mathtt{B}} \Gamma_i, (\nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}_i(\Delta \mathtt{W}^{(t,0)}_{i,r_i}) - \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}_i(\Delta \mathtt{W}^{(t)})) (\mathtt{A}^{(t)})^\top \Gamma_i \right\rangle \\
&\quad + \left\langle g^{(t)}_{\mathtt{B}} \Gamma_i, (\nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}_i(\Delta \mathtt{W}^{(t)}) - \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)})) (\mathtt{A}^{(t)})^\top \Gamma_i \right\rangle
\end{aligned}$$

Define $q_{\min} \triangleq \min_{l\in[L]} \sum_{i:r_{i,l}=r_l} p_i$, with $q_{\min} > 0$ since $r_l = \max_{i\in[N]} r_{i,l}$ guarantees that at least one client satisfies $r_l = r_{i,l}$ for each layer, and $p_i > 0$. Consequently, we ensure $q_{\min} > 0$. For client i with $r_{i,l} = r_l$, $\mathbf{\Gamma}_{i,l} = \mathbf{I}_{r_l}$. We derive two inequalities below:

$$\begin{aligned}
\sum_{i\in[N]} p_i \left\| \Gamma_i^\top g^{(t)}_{\mathtt{A}} \right\|_F^2 &= \sum_{l\in[L]} \sum_{i\in[N]} p_i \left\| \mathbf{\Gamma}_{i,l}^\top g^{(t)}_{\mathbf{A},l} \right\|_F^2 \geq \sum_{l\in[L]} \sum_{i:r_{i,l}=r_l} p_i \left\| \mathbf{\Gamma}_{i,l}^\top g^{(t)}_{\mathbf{A},l} \right\|_F^2 \geq \sum_{l\in[L]} \sum_{i:r_{i,l}=r_l} p_i \left\| g^{(t)}_{\mathbf{A},l} \right\|_F^2 \\
&\geq q_{\min} \sum_{l\in[L]} \left\| g^{(t)}_{\mathbf{A},l} \right\|_F^2 = q_{\min} \left\| g^{(t)}_{\mathtt{A}} \right\|_F^2,
\end{aligned}$$

and similarly, $\sum_{i\in[N]} p_i \left\| g^{(t)}_{\mathtt{B}} \Gamma_i \right\|_F^2 \geq q_{\min} \left\| g^{(t)}_{\mathtt{B}} \right\|_F^2$. By substituting the above terms into Part A, we have

$$\begin{aligned}
\mathbb{E}_t \left[\left\langle \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)}), \sum_{i\in[N]} p_i \delta^{(t)}_{i,r_i} \right\rangle \right] &= -\eta^{(t)} E \cdot \sum_{i\in[N]} p_i \left(\left\| \Gamma_i^\top g^{(t)}_{\mathtt{A}} \right\|_F^2 + \left\| g^{(t)}_{\mathtt{B}} \Gamma_i \right\|_F^2 \right) \\
&\quad - \eta^{(t)} E \cdot \sum_{i\in[N]} p_i \left\langle \Gamma_i^\top g^{(t)}_{\mathtt{A}}, (\mathtt{B}^{(t)} \Gamma_i)^\top (\nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}_i(\Delta \mathtt{W}^{(t,0)}_{i,r_i}) - \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}_i(\Delta \mathtt{W}^{(t)})) \right\rangle
\end{aligned}$$

$$
\begin{aligned}
&- \eta^{(t)} E \cdot \sum_{i\in[N]} p_i \Big\langle \Gamma_i^\top g_{\mathtt{A}}^{(t)}, (\mathtt{B}^{(t)}\Gamma_i)^\top (\nabla_{\Delta \mathtt{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}^{(t)}) - \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)})) \Big\rangle \\
&- \eta^{(t)} E \cdot \sum_{i\in[N]} p_i \Big\langle g_{\mathtt{B}}^{(t)}\Gamma_i, (\nabla_{\Delta \mathtt{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}_{i,r_i}^{(t,0)}) - \nabla_{\Delta \mathtt{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}^{(t)}))(\mathtt{A}^{(t)})^\top \Gamma_i \Big\rangle \\
&- \eta^{(t)} E \cdot \sum_{i\in[N]} p_i \Big\langle g_{\mathtt{B}}^{(t)}\Gamma_i, (\nabla_{\Delta \mathtt{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}^{(t)}) - \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)}))(\mathtt{A}^{(t)})^\top \Gamma_i \Big\rangle \\
&- \eta^{(t)} E \cdot \sum_{i\in[N]} p_i \Big\langle \Gamma_i^\top g_{\mathtt{A}}^{(t)}, D_{\mathtt{A},i,r_i}^{(t)} \Big\rangle - \eta^{(t)} E \cdot \sum_{i\in[N]} p_i \Big\langle g_{\mathtt{B}}^{(t)}\Gamma_i, D_{\mathtt{B},i,r_i}^{(t)} \Big\rangle \\
&+ (\eta^{(t)})^2 E^2 \cdot \Big\langle \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)}), \sum_{i\in[N]} p_i \mathbb{E}_t \Big[\bar{g}_{\mathtt{B},i,r_i}^{(t)} \bar{g}_{\mathtt{A},i,r_i}^{(t)} \Big] \Big\rangle \\
\overset{(a)}{=} &- \eta^{(t)} E \cdot \sum_{i\in[N]} p_i \Big(\Big\| \Gamma_i^\top g_{\mathtt{A}}^{(t)} \Big\|_F^2 + \Big\| g_{\mathtt{B}}^{(t)}\Gamma_i \Big\|_F^2 \Big) \\
&- \eta^{(t)} E \cdot \Big\langle g_{\mathtt{A}}^{(t)}, \sum_{i\in[N]} p_i \Gamma_i \Gamma_i^\top (\mathtt{B}^{(t)})^\top (\nabla_{\Delta \mathtt{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}_{i,r_i}^{(t,0)}) - \nabla_{\Delta \mathtt{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}^{(t)})) \Big\rangle \\
&- \eta^{(t)} E \cdot \Big\langle g_{\mathtt{A}}^{(t)}, \sum_{i\in[N]} p_i \Gamma_i \Gamma_i^\top (\mathtt{B}^{(t)})^\top (\nabla_{\Delta \mathtt{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}^{(t)}) - \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)})) \Big\rangle \\
&- \eta^{(t)} E \cdot \Big\langle g_{\mathtt{B}}^{(t)}, \sum_{i\in[N]} p_i (\nabla_{\Delta \mathtt{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}_{i,r_i}^{(t,0)}) - \nabla_{\Delta \mathtt{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}^{(t)}))(\mathtt{A}^{(t)})^\top \Gamma_i \Gamma_i^\top \Big\rangle \\
&- \eta^{(t)} E \cdot \Big\langle g_{\mathtt{B}}^{(t)}, \sum_{i\in[N]} p_i (\nabla_{\Delta \mathtt{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}^{(t)}) - \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)}))(\mathtt{A}^{(t)})^\top \Gamma_i \Gamma_i^\top \Big\rangle \\
&- \eta^{(t)} E \cdot \Big\langle g_{\mathtt{A}}^{(t)}, \sum_{i\in[N]} p_i \Gamma_i D_{\mathtt{A},i,r_i}^{(t)} \Big\rangle - \eta^{(t)} E \cdot \Big\langle g_{\mathtt{B}}^{(t)}, \sum_{i\in[N]} p_i D_{\mathtt{B},i,r_i}^{(t)} \Gamma_i^\top \Big\rangle \\
&+ (\eta^{(t)})^2 E^2 \cdot \Big\langle \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)}), \sum_{i\in[N]} p_i \mathbb{E}_t \Big[\bar{g}_{\mathtt{B},i,r_i}^{(t)} \bar{g}_{\mathtt{A},i,r_i}^{(t)} \Big] \Big\rangle \\
\overset{(b)}{\leq} &- \eta^{(t)} E \cdot q_{\min} (\Big\| g_{\mathtt{A}}^{(t)} \Big\|_F^2 + \Big\| g_{\mathtt{B}}^{(t)} \Big\|_F^2) \\
&+ \eta^{(t)} E \Big(\frac{q_{\min}}{6} \Big\| g_{\mathtt{A}}^{(t)} \Big\|_F^2 + \frac{3}{2q_{\min}} \Big\| \sum_{i\in[N]} p_i \Gamma_i \Gamma_i^\top (\mathtt{B}^{(t)})^\top (\nabla_{\Delta \mathtt{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}_{i,r_i}^{(t,0)}) - \nabla_{\Delta \mathtt{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}^{(t)})) \Big\|_F^2 \Big) \\
&+ \eta^{(t)} E \Big(\frac{q_{\min}}{6} \Big\| g_{\mathtt{A}}^{(t)} \Big\|_F^2 + \frac{3}{2q_{\min}} \Big\| \sum_{i\in[N]} p_i \Gamma_i \Gamma_i^\top (\mathtt{B}^{(t)})^\top (\nabla_{\Delta \mathtt{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}^{(t)}) - \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)})) \Big\|_F^2 \Big) \\
&+ \eta^{(t)} E \Big(\frac{q_{\min}}{6} \Big\| g_{\mathtt{B}}^{(t)} \Big\|_F^2 + \frac{3}{2q_{\min}} \Big\| \sum_{i\in[N]} p_i (\nabla_{\Delta \mathtt{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}_{i,r_i}^{(t,0)}) - \nabla_{\Delta \mathtt{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}^{(t)}))(\mathtt{A}^{(t)})^\top \Gamma_i \Gamma_i^\top \Big\|_F^2 \Big) \\
&+ \eta^{(t)} E \Big(\frac{q_{\min}}{6} \Big\| g_{\mathtt{B}}^{(t)} \Big\|_F^2 + \frac{3}{2q_{\min}} \Big\| \sum_{i\in[N]} p_i (\nabla_{\Delta \mathtt{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}^{(t)}) - \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)}))(\mathtt{A}^{(t)})^\top \Gamma_i \Gamma_i^\top \Big\|_F^2 \Big) \\
&+ \eta^{(t)} E \Big(\frac{q_{\min}}{6} \Big\| g_{\mathtt{A}}^{(t)} \Big\|_F^2 + \frac{3}{2q_{\min}} \Big\| \sum_{i\in[N]} p_i \Gamma_i D_{\mathtt{A},i,r_i}^{(t)} \Big\|_F^2 \Big) \\
&+ \eta^{(t)} E \Big(\frac{q_{\min}}{6} \Big\| g_{\mathtt{B}}^{(t)} \Big\|_F^2 + \frac{3}{2q_{\min}} \Big\| \sum_{i\in[N]} p_i D_{\mathtt{B},i,r_i}^{(t)} \Gamma_i^\top \Big\|_F^2 \Big)
\end{aligned}
$$

$$
\begin{aligned}
&+ (\eta^{(t)})^2 E^2 \Big\| \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)}) \Big\|_F \Big\| \sum_{i\in[N]} p_i \mathbb{E}_t \Big[\bar{g}^{(t)}_{\mathtt{B},i,r_i} \bar{g}^{(t)}_{\mathtt{A},i,r_i} \Big] \Big\|_F \\
&\overset{(c)}{\leq} -\eta^{(t)} E \cdot \frac{q_{\min}}{2} \Big(\Big\| g^{(t)}_{\mathtt{A}} \Big\|_F^2 + \Big\| g^{(t)}_{\mathtt{B}} \Big\|_F^2 \Big) \\
&+ \eta^{(t)} E \cdot \frac{3}{2q_{\min}} \sum_{i\in[N]} p_i \Big\| \Gamma_i \Gamma_i^\top (\mathtt{B}^{(t)})^\top (\nabla_{\Delta \mathtt{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}^{(t,0)}_{i,r_i}) - \nabla_{\Delta \mathtt{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}^{(t)})) \Big\|_F^2 \\
&+ \eta^{(t)} E \cdot \frac{3}{2q_{\min}} \sum_{i\in[N]} p_i \Big\| \Gamma_i \Gamma_i^\top (\mathtt{B}^{(t)})^\top (\nabla_{\Delta \mathtt{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}^{(t)}) - \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)})) \Big\|_F^2 \\
&+ \eta^{(t)} E \cdot \frac{3}{2q_{\min}} \sum_{i\in[N]} p_i \Big\| (\nabla_{\Delta \mathtt{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}^{(t,0)}_{i,r_i}) - \nabla_{\Delta \mathtt{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}^{(t)})) (\mathtt{A}^{(t)})^\top \Gamma_i \Gamma_i^\top \Big\|_F^2 \\
&+ \eta^{(t)} E \cdot \frac{3}{2q_{\min}} \sum_{i\in[N]} p_i \Big\| (\nabla_{\Delta \mathtt{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}^{(t)}) - \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)})) (\mathtt{A}^{(t)})^\top \Gamma_i \Gamma_i^\top \Big\|_F^2 \\
&+ \eta^{(t)} E \cdot \frac{3}{2q_{\min}} \sum_{i\in[N]} p_i \Big\| \Gamma_i D^{(t)}_{\mathtt{A},i,r_i} \Big\|_F^2 + \eta^{(t)} E \cdot \frac{3}{2q_{\min}} \sum_{i\in[N]} p_i \Big\| D^{(t)}_{\mathtt{B},i,r_i} \Gamma_i^\top \Big\|_F^2 \\
&+ (\eta^{(t)})^2 E^2 \cdot \Big\| \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)}) \Big\|_F \cdot \sum_{i\in[N]} p_i \Big\| \mathbb{E}_t \Big[\bar{g}^{(t)}_{\mathtt{B},i,r_i} \bar{g}^{(t)}_{\mathtt{A},i,r_i} \Big] \Big\|_F \\
&\overset{(d)}{\leq} -\eta^{(t)} E \cdot \frac{q_{\min}}{2} \Big(\Big\| g^{(t)}_{\mathtt{A}} \Big\|_F^2 + \Big\| g^{(t)}_{\mathtt{B}} \Big\|_F^2 \Big) \\
&+ \eta^{(t)} E \cdot \frac{3}{2q_{\min}} (M_{\mathbf{B}}^2 + M_{\mathbf{A}}^2) \mu^2 \sum_{i\in[N]} p_i \Big\| \Delta \mathtt{W}^{(t,0)}_{i,r_i} - \Delta \mathtt{W}^{(t)} \Big\|_F^2 \\
&+ \eta^{(t)} E \cdot \frac{3}{2q_{\min}} (M_{\mathbf{B}}^2 + M_{\mathbf{A}}^2) \sum_{i\in[N]} p_i \Big\| \nabla_{\Delta \mathtt{W}} \mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}^{(t)}) - \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)}) \Big\|_F^2 \\
&+ \eta^{(t)} E \cdot \frac{3}{2q_{\min}} \sum_{i\in[N]} p_i \Big\| D^{(t)}_{\mathtt{A},i,r_i} \Big\|_F^2 + \eta^{(t)} E \cdot \frac{3}{2q_{\min}} \sum_{i\in[N]} p_i \Big\| D^{(t)}_{\mathtt{B},i,r_i} \Big\|_F^2 \\
&+ (\eta^{(t)})^2 E^2 \cdot \Big\| \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)}) \Big\|_F \cdot \sum_{i\in[N]} p_i \Big\| \mathbb{E}_t \Big[\bar{g}^{(t)}_{\mathtt{B},i,r_i} \bar{g}^{(t)}_{\mathtt{A},i,r_i} \Big] \Big\|_F \\
&\overset{(e)}{\leq} -\eta^{(t)} E \cdot \frac{q_{\min}}{2} \Big(\Big\| g^{(t)}_{\mathtt{A}} \Big\|_F^2 + \Big\| g^{(t)}_{\mathtt{B}} \Big\|_F^2 \Big) \\
&+ \eta^{(t)} E \cdot \frac{3}{2q_{\min}} (M_{\mathbf{B}}^2 + M_{\mathbf{A}}^2) \mu^2 \sum_{i\in[N]} p_i \Big\| \Delta \mathtt{W}^{(t,0)}_{i,r_i} - \Delta \mathtt{W}^{(t)} \Big\|_F^2 \\
&+ \eta^{(t)} E \cdot \frac{3}{2q_{\min}} (M_{\mathbf{B}}^2 + M_{\mathbf{A}}^2) \zeta^2 \\
&+ (\eta^{(t)})^3 \frac{(E-1)E(2E-1)}{3} \cdot \frac{3}{2q_{\min}} \chi^2 \Big(\chi^2 (M_{\mathbf{A}}^2 + M_{\mathbf{B}}^2) + 2\mu^2 (M_{\mathbf{A}}^2 + M_{\mathbf{B}}^2)(M_{\mathbf{B}}^4 + M_{\mathbf{A}}^4) \Big) \\
&+ (\eta^{(t)})^2 E^2 \cdot \chi \cdot M_{\mathbf{A}} M_{\mathbf{B}} \kappa^2,
\end{aligned}
$$

where (a) holds since the definition of the tuple inner product and the cyclic property of the trace, the first term in (b) holds since $\sum_{i\in[N]} p_i \Big\| \Gamma_i^\top g^{(t)}_{\mathtt{A}} \Big\|_F^2 \geq q_{\min} \Big\| g^{(t)}_{\mathtt{A}} \Big\|_F^2$ and $\sum_{i\in[N]} p_i \Big\| g^{(t)}_{\mathtt{B}} \Gamma_i \Big\|_F^2 \geq q_{\min} \Big\| g^{(t)}_{\mathtt{B}} \Big\|_F^2$, the middle six terms hold due to Young's inequality, i.e., $\langle X, Y \rangle \leq \frac{\lambda}{2} \|X\|_F^2 + \frac{1}{2\lambda} \|Y\|_F^2$, and choosing $\lambda = \frac{q_{\min}}{3}$, and the last term holds since $\langle X, Y \rangle \leq |\langle X, Y \rangle| \leq \|X\|_F \|Y\|_F$, (c) is obtained by rearranging the factors of $\Big\| g^{(t)}_{\mathtt{A}} \Big\|_F^2$ and $\Big\| g^{(t)}_{\mathtt{B}} \Big\|_F^2$, using Jensen's inequality, (d) holds by using $\|XY\|_F \leq \|X\|_2 \|Y\|_F$, $\mathbf{\Gamma}_{i,l}^\top \mathbf{\Gamma}_{i,l} = \mathbf{I}_{r_{i,l}}$, Assumption 4, and $\|\mathbf{\Gamma}_{i,l} \mathbf{\Gamma}_{i,l}^\top\|_2 = 1$, (e) is obtained

by using Assumption 3, Lemma 4 Lemma 5, and

$$\begin{aligned}\left\|\nabla_{\Delta \mathtt{W}}\mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)})\right\|_F^2 &= \left\|\sum_{i\in[N]} p_i \nabla_{\Delta \mathtt{W}}\mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}^{(t)})\right\|_F^2 \\ &\le \sum_{i\in[N]} p_i \left\|\nabla_{\Delta \mathtt{W}}\mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}^{(t)})\right\|_F^2 \\ &= \sum_{i\in[N]} p_i \left\|\mathbb{E}_t\left[\nabla_{\Delta \mathtt{W}}\mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}^{(t)};\xi_i)\right]\right\|_F^2 \\ &\le \sum_{i\in[N]} p_i \mathbb{E}_t\left[\left\|\nabla_{\Delta \mathtt{W}}\mathcal{L}_i^{\text{full}}(\Delta \mathtt{W}^{(t)};\xi_i)\right\|_F^2\right] \\ &\le \chi^2,\end{aligned}$$

We now derive the bound for $\sum_{i\in[N]} p_i \left\|\Delta \mathtt{W}_{i,r_i}^{(t,0)} - \Delta \mathtt{W}^{(t)}\right\|_F^2$. For layer $l \in [L]$:

$$\begin{aligned}&\Delta \mathbf{W}_{i,l}^{(t,0)} - \Delta \mathbf{W}_l^{(t)} = \mathbf{B}_{i,l}^{(t,0)}\mathbf{A}_{i,l}^{(t,0)} - \mathbf{B}_l^{(t)}\mathbf{A}_l^{(t)} = \mathbf{B}_l^{(t)}\mathbf{\Gamma}_{i,l}(\mathbf{\Gamma}_{i,l})^\top \mathbf{A}_l^{(t)} - \mathbf{B}_l^{(t)}\mathbf{A}_l^{(t)} \\ &= -\mathbf{B}_l^{(t)}(\mathbf{I}_{r_l} - \mathbf{\Gamma}_{i,l}(\mathbf{\Gamma}_{i,l})^\top)\mathbf{A}_l^{(t)} = -\widehat{\mathbf{B}}_l^{(t)}\mathbf{U}_l^{(t)}(\mathbf{\Sigma}_l^{(t)})^{\frac{1}{2}}(\mathbf{I}_{r_l} - \mathbf{\Gamma}_{i,l}(\mathbf{\Gamma}_{i,l})^\top)(\mathbf{\Sigma}_l^{(t)})^{\frac{1}{2}}(\mathbf{V}_l^{(t)})^\top \\ &= -\widehat{\mathbf{B}}_l^{(t)}\mathbf{U}_l^{(t)}(\mathbf{I}_{r_l} - \mathbf{\Gamma}_{i,l}(\mathbf{\Gamma}_{i,l})^\top)\mathbf{\Sigma}_l^{(t)}(\mathbf{V}_l^{(t)})^\top\end{aligned}$$

Consequently,

$$\begin{aligned}\left\|\Delta \mathbf{W}_{i,l}^{(t,0)} - \Delta \mathbf{W}_l^{(t)}\right\|_F^2 &= \left\| - \widehat{\mathbf{B}}_l^{(t)}\mathbf{U}_{l,\text{tail}}^{(t)}\mathbf{\Sigma}_{l,\text{tail}}^{(t)}(\mathbf{V}_{l,\text{tail}}^{(t)})^\top\right\|_F^2 \\ &= \text{tr}\Big(\mathbf{V}_{l,\text{tail}}^{(t)}(\mathbf{\Sigma}_{l,\text{tail}}^{(t)})^\top(\mathbf{U}_{l,\text{tail}}^{(t)})^\top(\widehat{\mathbf{B}}_l^{(t)})^\top\widehat{\mathbf{B}}_l^{(t)}\mathbf{U}_{l,\text{tail}}^{(t)}\mathbf{\Sigma}_{l,\text{tail}}^{(t)}(\mathbf{V}_{l,\text{tail}}^{(t)})^\top\Big) \\ &\overset{(a)}{=} \text{tr}\Big(\mathbf{V}_{l,\text{tail}}^{(t)}(\mathbf{\Sigma}_{l,\text{tail}}^{(t)})^2(\mathbf{V}_{l,\text{tail}}^{(t)})^\top\Big) = \text{tr}\Big((\mathbf{\Sigma}_{l,\text{tail}}^{(t)})^2(\mathbf{V}_{l,\text{tail}}^{(t)})^\top\mathbf{V}_{l,\text{tail}}^{(t)}\Big) \\ &= \text{tr}\Big((\mathbf{\Sigma}_{l,\text{tail}}^{(t)})^2\Big) = \sum_{j=r_{i,l}+1}^{r_l}\left(\sigma_{l,j}^{(t)}\right)^2,\end{aligned}$$

where the subscript "tail" represents the columns/rows that correspond to rank $r_{i,l} < j \le r_l$, $\sigma_{l,j}^{(t)}$ is the j-th singular value of $\Delta \mathbf{W}_l^{(t)}$, and (a) holds since both $\widehat{\mathbf{B}}_l^{(t)}$ and $\mathbf{U}_{l,\text{tail}}^{(t)}$ have orthonormal columns. Hence, $\sum_{i\in[N]} p_i \left\|\Delta \mathtt{W}_{i,r_i}^{(t,0)} - \Delta \mathtt{W}^{(t)}\right\|_F^2 = \sum_{i\in[N]} p_i \sum_{l\in[L]}\sum_{j=r_{i,l}+1}^{r_l}\left(\sigma_{l,j}^{(t)}\right)^2$.

For Part B, we have

$$\mathbb{E}_t\left[\left\langle \nabla_{\Delta \mathtt{W}}\mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)}), \mathcal{E}_{\text{sketch}}^{(t+1)}\right\rangle\right] \le \left\|\nabla_{\Delta \mathtt{W}}\mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)})\right\|_F \mathbb{E}_t\left\|\mathcal{E}_{\text{sketch}}^{(t+1)}\right\|_F \le \chi \mathbb{E}_t\left\|\mathcal{E}_{\text{sketch}}^{(t+1)}\right\|_F.$$

By combining the previous bounds on Part A and Part B, Term 1 is bounded as:

$$\begin{aligned}&\mathbb{E}_t\left[\left\langle \nabla_{\Delta \mathtt{W}}\mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)}), \Delta \mathtt{W}_{\text{SEFORA}}^{(t+1)} - \Delta \mathtt{W}^{(t)}\right\rangle\right] \le -\eta^{(t)}E \cdot \frac{q_{\min}}{2}\left(\left\|g_{\mathtt{A}}^{(t)}\right\|_F^2 + \left\|g_{\mathtt{B}}^{(t)}\right\|_F^2\right) \\ &+ \eta^{(t)}E \cdot \frac{3}{2q_{\min}}(M_{\mathbf{B}}^2 + M_{\mathbf{A}}^2)\mu^2 \sum_{i\in[N]} p_i \sum_{l\in[L]}\sum_{j=r_{i,l}+1}^{r_l}\left(\sigma_{l,j}^{(t)}\right)^2 \\ &+ \eta^{(t)}E \cdot \frac{3}{2q_{\min}}(M_{\mathbf{B}}^2 + M_{\mathbf{A}}^2)\zeta^2 \\ &+ (\eta^{(t)})^3\frac{(E-1)E(2E-1)}{3} \cdot \frac{3}{2q_{\min}}\chi^2\Big(\chi^2(M_{\mathbf{A}}^2 + M_{\mathbf{B}}^2) + 2\mu^2(M_{\mathbf{A}}^2 + M_{\mathbf{B}}^2)(M_{\mathbf{B}}^4 + M_{\mathbf{A}}^4)\Big) \\ &+ (\eta^{(t)})^2E^2 \cdot \chi \cdot M_{\mathbf{A}}M_{\mathbf{B}}\kappa^2 + \chi \cdot \mathbb{E}_t\left\|\mathcal{E}_{\text{sketch}}^{(t+1)}\right\|_F.\end{aligned}$$

For Term 2:

$$
\begin{aligned}
&\mathbb{E}_t\left[\|\Delta \mathtt{W}^{(t+1)}_{\textsc{SeFoRA}} - \Delta \mathtt{W}^{(t)}\|_F^2\right] \\
&= \mathbb{E}_t\left[\Big\| \sum_{i\in[N]} p_i \Big(-\eta^{(t)} E \mathtt{B}^{(t)}\Gamma_i \bar{g}^{(t)}_{\mathtt{A},i,r_i} - \eta^{(t)} E \bar{g}^{(t)}_{\mathtt{B},i,r_i} (\Gamma_i)^\top \mathtt{A}^{(t)} + (\eta^{(t)})^2 E^2 \bar{g}^{(t)}_{\mathtt{B},i,r_i} \bar{g}^{(t)}_{\mathtt{A},i,r_i}\Big) + \mathcal{E}^{(t+1)}_{\text{sketch}} \Big\|_F^2\right] \\
&\overset{(a)}{\leq} 4(\eta^{(t)})^2 E^2 \cdot \sum_{i\in[N]} p_i \mathbb{E}_t\left[\Big\| \mathtt{B}^{(t)}\Gamma_i \cdot \bar{g}^{(t)}_{\mathtt{A},i,r_i}\Big\|_F^2\right] + 4(\eta^{(t)})^2 E^2 \cdot \sum_{i\in[N]} p_i \mathbb{E}_t\left[\Big\| \bar{g}^{(t)}_{\mathtt{B},i,r_i} \cdot (\Gamma_i)^\top \mathtt{A}^{(t)}\Big\|_F^2\right] \\
&\quad + 4(\eta^{(t)})^4 E^4 \sum_{i\in[N]} p_i \mathbb{E}_t\left[\Big\| \bar{g}^{(t)}_{\mathtt{B},i,r_i} \bar{g}^{(t)}_{\mathtt{A},i,r_i}\Big\|_F^2\right] + 4\mathbb{E}_t\left[\Big\|\mathcal{E}^{(t+1)}_{\text{sketch}}\Big\|_F^2\right] \\
&\overset{(b)}{\leq} 4(\eta^{(t)})^2 E^2 \sum_{i\in[N]} p_i \mathbb{E}_t\left[\Big\| \mathtt{B}^{(t)}\Gamma_i\Big\|_2^2 \Big\| \bar{g}^{(t)}_{\mathtt{A},i,r_i}\Big\|_F^2\right] + 4(\eta^{(t)})^2 E^2 \sum_{i\in[N]} p_i \mathbb{E}_t\left[\Big\| \bar{g}^{(t)}_{\mathtt{B},i,r_i}\Big\|_F^2 \Big\| (\Gamma_i)^\top \mathtt{A}^{(t)}\Big\|_2^2\right] \\
&\quad + 4(\eta^{(t)})^4 E^4 \sum_{i\in[N]} p_i \mathbb{E}_t\left[\Big\| \bar{g}^{(t)}_{\mathtt{B},i,r_i} \bar{g}^{(t)}_{\mathtt{A},i,r_i}\Big\|_F^2\right] + 4\mathbb{E}_t\left[\Big\|\mathcal{E}^{(t+1)}_{\text{sketch}}\Big\|_F^2\right] \\
&\overset{(c)}{\leq} 4(\eta^{(t)})^2 E^2 \cdot (M_{\mathbf{A}}^4 + M_{\mathbf{B}}^4)\chi^2 + 4(\eta^{(t)})^4 E^4 M_{\mathbf{A}}^2 M_{\mathbf{B}}^2 \kappa^4 + 4\mathbb{E}_t\left[\Big\|\mathcal{E}^{(t+1)}_{\text{sketch}}\Big\|_F^2\right]
\end{aligned}
$$

where (a) holds by using $\|W+X+Y+Z\|_F^2 \leq 4(\|W\|_F^2 + \|X\|_F^2 + \|Y\|_F^2 + \|Z\|_F^2)$ and Jensen's inequality, (b) uses $\|XY\|_F \leq \|X\|_2 \|Y\|_F$, (c) holds since we have $\left\|\mathbf{B}_l^{(t)}\mathbf{\Gamma}_{i,l}\right\|_2 \leq \left\|\mathbf{B}_l^{(t)}\right\|_2 \left\|\mathbf{\Gamma}_{i,l}\right\|_2 \leq M_{\mathbf{B}}$, Lemma 5 and Lemma 6.

By combining the bounds on Term 1 and Term 2 into Equation (2), we have

$$
\begin{aligned}
&\mathbb{E}_t\left[\mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t+1)}_{\textsc{SeFoRA}})\right] - \mathbb{E}_t\left[\mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)})\right] \\
&\leq \mathbb{E}_t\left[\left\langle \nabla_{\Delta \mathtt{W}} \mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)}), \Delta \mathtt{W}^{(t+1)}_{\textsc{SeFoRA}} - \Delta \mathtt{W}^{(t)}\right\rangle\right] + \frac{\mu}{2}\mathbb{E}_t\left[\|\Delta \mathtt{W}^{(t+1)}_{\textsc{SeFoRA}} - \Delta \mathtt{W}^{(t)}\|_F^2\right] \\
&\leq -\eta^{(t)} E \cdot \frac{q_{\min}}{2}\Big(\Big\|g_{\mathtt{A}}^{(t)}\Big\|_F^2 + \Big\|g_{\mathtt{B}}^{(t)}\Big\|_F^2\Big) \\
&\quad + \eta^{(t)} E \cdot \frac{3}{2q_{\min}}(M_{\mathbf{B}}^2 + M_{\mathbf{A}}^2)\Big(\mu^2 \sum_{i\in[N]} p_i \sum_{l\in[L]} \sum_{j=r_{i,l}+1}^{r_l} \big(\sigma_{l,j}^{(t)}\big)^2 + \zeta^2\Big) \\
&\quad + (\eta^{(t)})^2 E^2 \cdot \chi \cdot M_{\mathbf{A}} M_{\mathbf{B}} \kappa^2 + (\eta^{(t)})^2 E^2 \cdot 2\mu(M_{\mathbf{A}}^4 + M_{\mathbf{B}}^4)\chi^2 \\
&\quad + (\eta^{(t)})^3 E^3 \frac{1}{q_{\min}} \chi^2 \Big(\chi^2(M_{\mathbf{A}}^2 + M_{\mathbf{B}}^2) + 2\mu^2(M_{\mathbf{A}}^2 + M_{\mathbf{B}}^2)(M_{\mathbf{B}}^4 + M_{\mathbf{A}}^4)\Big) \\
&\quad + (\eta^{(t)})^4 E^4 \cdot 2\mu M_{\mathbf{A}}^2 M_{\mathbf{B}}^2 \kappa^4 + \chi \cdot \mathbb{E}_t\Big\|\mathcal{E}^{(t+1)}_{\text{sketch}}\Big\|_F + 2\mu\mathbb{E}_t\left[\Big\|\mathcal{E}^{(t+1)}_{\text{sketch}}\Big\|_F^2\right].
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\Big\|g_{\mathtt{A}}^{(t)}\Big\|_F^2 + \Big\|g_{\mathtt{B}}^{(t)}\Big\|_F^2 &\leq \frac{2}{\eta^{(t)} E q_{\min}}(\mathbb{E}_t\left[\mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t)})\right] - \mathbb{E}_t\left[\mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(t+1)}_{\textsc{SeFoRA}})\right]) \\
&\quad + \frac{6}{2q_{\min}^2}(M_{\mathbf{B}}^2 + M_{\mathbf{A}}^2)\Big(\mu^2 \sum_{i\in[N]} p_i \sum_{l\in[L]} \sum_{j=r_{i,l}+1}^{r_l} \big(\sigma_{l,j}^{(t)}\big)^2 + \zeta^2\Big) \\
&\quad + (\eta^{(t)}) E \cdot \frac{2}{q_{\min}}\Big(\chi M_{\mathbf{A}} M_{\mathbf{B}} \kappa^2 + 2\mu(M_{\mathbf{A}}^4 + M_{\mathbf{B}}^4)\chi^2\Big) \\
&\quad + (\eta^{(t)})^2 E^2 \cdot \frac{2}{q_{\min}^2} \chi^2 \Big(\chi^2(M_{\mathbf{A}}^2 + M_{\mathbf{B}}^2) + 2\mu^2(M_{\mathbf{A}}^2 + M_{\mathbf{B}}^2)(M_{\mathbf{B}}^4 + M_{\mathbf{A}}^4)\Big)
\end{aligned}
$$

$$+ (\eta^{(t)})^3 E^3 \cdot \frac{4}{q_{\min}} \mu M_{\mathbf{A}}^2 M_{\mathbf{B}}^2 \kappa^4 + \frac{2}{\eta^{(t)} E q_{\min}} \left(\chi \cdot \mathbb{E}_t \left\| \mathcal{E}_{\text{sketch}}^{(t+1)} \right\|_F + 2\mu \mathbb{E}_t \left\| \mathcal{E}_{\text{sketch}}^{(t+1)} \right\|_F^2 \right).$$

By taking constant learning rate $\eta^{(t)} = \eta$, taking total expectation $\mathbb{E}[\cdot]$ on both sides, summing this inequality from $t = 0$ to $T-1$, computing the average, and defining $\Delta_0 = \mathbb{E}\left[\mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(0)})\right] - \mathcal{L}^{\text{full},\star}$, we have

$$\begin{aligned}
&\frac{1}{T} \sum_{t=0}^{T-1} \left(\mathbb{E}\left[\left\| \nabla_{\mathtt{A}} \mathcal{L}^{\text{lora}}(\mathtt{B}^{(t)}, \mathtt{A}^{(t)}) \right\|_F^2 \right] + \mathbb{E}\left[\left\| \nabla_{\mathtt{B}} \mathcal{L}^{\text{lora}}(\mathtt{B}^{(t)}, \mathtt{A}^{(t)}) \right\|_F^2 \right] \right) \\
&\leq \frac{2\Delta_0}{\eta E T q_{\min}} + \frac{1}{T} \sum_{t=0}^{T-1} \frac{2}{\eta E q_{\min}} \mathbb{E}\left[\chi \cdot \mathbb{E}_t \left\| \mathcal{E}_{\text{sketch}}^{(t+1)} \right\|_F + 2\mu \mathbb{E}_t \left\| \mathcal{E}_{\text{sketch}}^{(t+1)} \right\|_F^2 \right] \\
&\quad + \frac{6}{2q_{\min}^2} (M_{\mathbf{B}}^2 + M_{\mathbf{A}}^2) \left(\mu^2 \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\Big[\sum_{i \in [N]} p_i \sum_{l \in [L]} \sum_{j=r_{i,l}+1}^{r_l} \left(\sigma_{l,j}^{(t)}\right)^2 \Big] + \zeta^2 \right) \\
&\quad + \eta E \cdot \frac{2}{q_{\min}} \left(\chi M_{\mathbf{A}} M_{\mathbf{B}} \kappa^2 + 2\mu (M_{\mathbf{A}}^4 + M_{\mathbf{B}}^4) \chi^2 \right) \\
&\quad + \eta^2 E^2 \frac{2}{q_{\min}^2} \chi^2 \left(\chi^2 (M_{\mathbf{A}}^2 + M_{\mathbf{B}}^2) + 2\mu^2 (M_{\mathbf{A}}^2 + M_{\mathbf{B}}^2)(M_{\mathbf{B}}^4 + M_{\mathbf{A}}^4) \right) + \eta^3 E^3 \frac{4}{q_{\min}} \mu M_{\mathbf{A}}^2 M_{\mathbf{B}}^2 \kappa^4.
\end{aligned}$$

Recall that $\mathcal{E}_{\text{sketch}}^{(t+1)} = \Delta \mathtt{W}_{\text{SEFORA}}^{(t+1)} - \mathtt{V}^{(t+1)}$. Let $C_1 = M_{\mathbf{B}}^2 + M_{\mathbf{A}}^2$, $C_2 = \chi M_{\mathbf{A}} M_{\mathbf{B}} \kappa^2 + 2\mu(M_{\mathbf{A}}^4 + M_{\mathbf{B}}^4)\chi^2$, $C_3 = \chi^2 \left(\chi^2 (M_{\mathbf{A}}^2 + M_{\mathbf{B}}^2) + 2\mu^2 (M_{\mathbf{A}}^2 + M_{\mathbf{B}}^2)(M_{\mathbf{B}}^4 + M_{\mathbf{A}}^4) \right)$, $C_4 = \mu M_{\mathbf{A}}^2 M_{\mathbf{B}}^2 \kappa^4$, the sketching floor $S_1^{(t+1)} = \chi \cdot \mathbb{E}_t \left\| \mathcal{E}_{\text{sketch}}^{(t+1)} \right\|_F + 2\mu \mathbb{E}_t \left\| \mathcal{E}_{\text{sketch}}^{(t+1)} \right\|_F^2$, $\overline{S_1} = \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[S_1^{(t+1)} \right]$, and $\Delta_0 = \mathbb{E}\left[\mathcal{L}^{\text{full}}(\Delta \mathtt{W}^{(0)})\right] - \mathcal{L}^{\text{full},\star}$. SEFORA satisfies:

$$\begin{aligned}
&\frac{1}{T} \sum_{t=0}^{T-1} \left(\mathbb{E}\left[\left\| \nabla_{\mathtt{A}} \mathcal{L}^{\text{lora}}(\mathtt{B}^{(t)}, \mathtt{A}^{(t)}) \right\|_F^2 \right] + \mathbb{E}\left[\left\| \nabla_{\mathtt{B}} \mathcal{L}^{\text{lora}}(\mathtt{B}^{(t)}, \mathtt{A}^{(t)}) \right\|_F^2 \right] \right) \\
&\leq \frac{2\Delta_0}{\eta E T q_{\min}} + \frac{2}{\eta E q_{\min}} \overline{S_1} + \eta E \cdot \frac{2C_2}{q_{\min}} + \eta^2 E^2 \cdot \frac{2C_3}{q_{\min}^2} + \eta^3 E^3 \cdot \frac{4C_4}{q_{\min}} \\
&\quad + \frac{6C_1}{2q_{\min}^2} \left(\mu^2 \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\Big[\sum_{i \in [N]} p_i \sum_{l \in [L]} \sum_{j=r_{i,l}+1}^{r_l} \left(\sigma_{l,j}^{(t)}\right)^2 \Big] + \zeta^2 \right).
\end{aligned}$$

This concludes the proof of Theorem 1

□

E.2 Proof of Corollary 1

Proof. Recall that $\mathcal{E}_{\text{sketch}}^{(t+1)} = \Delta \mathtt{W}_{\text{SEFORA}}^{(t+1)} - \mathtt{V}^{(t+1)}$, where $\mathtt{V}^{(t+1)} \triangleq \Delta \mathtt{W}^{(t)} + \sum_{i \in [N]} p_i (\Delta \mathtt{W}_{i,r_i}^{(t,E)} - \Delta \mathtt{W}_{i,r_i}^{(t,0)})$. According to Lemma 1, the sketching error is bounded as:

$$\begin{aligned}
\mathbb{E}_t \|\mathcal{E}_{\text{sketch}}^{(t+1)}\|_F^2 &= \sum_{l=1}^{L} \mathbb{E}_t \|\mathcal{E}_{\text{sketch},l}^{(t+1)}\|_F^2 \\
&\leq \sum_{l=1}^{L} (1 + \frac{r_l}{k_l - r_l - 1}) \cdot \min_{\varrho_l' < r_l - 1} (1 + \frac{\varrho_l'}{r_l - \varrho_l' - 1}) \cdot \mathbb{E}_t \left[\tau_{\varrho_l'+1}^2 (\mathbf{V}_l^{(t+1)}) \right] \\
&\leq \sum_{l=1}^{L} c_l \cdot \mathbb{E}_t \left[\tau_{\varrho_l+1}^2 (\mathbf{V}_l^{(t+1)}) \right],
\end{aligned}$$

where the second inequality holds by evaluating at the fixed ϱ_l from Assumption 5, and $c_l \triangleq (1 + \frac{r_l}{k_l - r_l - 1}) \cdot (1 + \frac{\varrho_l}{r_l - \varrho_l - 1})$.

Define $\epsilon^2 \triangleq \sum_{l=1}^{L} c_l \cdot \overline{\tau_l}^2$. Averaging over the T communication rounds and invoking Assumption 5, we obtain

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\mathcal{E}_{\text{sketch}}^{(t+1)}\|_F^2\right] \le \frac{1}{T}\sum_{t=0}^{T-1}\sum_{l=1}^{L} c_l \cdot \mathbb{E}_t\left[\tau_{\varrho_l+1}^2(\mathbf{V}_l^{(t+1)})\right] \le \sum_{l=1}^{L} c_l \cdot \overline{\tau_l}^2 = \epsilon^2.$$

Thus, the sketching floor is

$$\begin{aligned}
\overline{S_1} &= \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[S_1^{(t+1)}\right] = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\chi\cdot\mathbb{E}_t\left\|\mathcal{E}_{\text{sketch}}^{(t+1)}\right\|_F + 2\mu\mathbb{E}_t\left\|\mathcal{E}_{\text{sketch}}^{(t+1)}\right\|_F^2\right] \\
&\le \chi\cdot\frac{1}{T}\sum_{t=0}^{T-1}\sqrt{\mathbb{E}\left[\left\|\mathcal{E}_{\text{sketch}}^{(t+1)}\right\|_F^2\right]} + 2\mu\cdot\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\mathcal{E}_{\text{sketch}}^{(t+1)}\|_F^2\right] \\
&\le \chi\cdot\sqrt{\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\mathcal{E}_{\text{sketch}}^{(t+1)}\right\|_F^2\right]} + 2\mu\cdot\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\|\mathcal{E}_{\text{sketch}}^{(t+1)}\|_F^2\right] \\
&\le \chi\epsilon + 2\mu\epsilon^2,
\end{aligned}$$

where the inequality holds due to Jensen's inequality.

By substituting this bound into the result of Theorem 1, we obtain

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\left(\mathbb{E}\left[\left\|\nabla_{\mathtt{A}}\mathcal{L}^{\text{lora}}(\mathtt{B}^{(t)},\mathtt{A}^{(t)})\right\|_F^2\right] + \mathbb{E}\left[\left\|\nabla_{\mathtt{B}}\mathcal{L}^{\text{lora}}(\mathtt{B}^{(t)},\mathtt{A}^{(t)})\right\|_F^2\right]\right) \\
&\le \frac{2\Delta_0}{\eta E T q_{\min}} + \frac{2}{\eta E q_{\min}}(\chi\epsilon + 2\mu\epsilon^2) + \eta E\cdot\frac{2C_2}{q_{\min}} + \eta^2E^2\cdot\frac{2C_3}{q_{\min}^2} + \eta^3E^3\cdot\frac{4C_4}{q_{\min}} \\
&\quad + \frac{6C_1}{2q_{\min}^2}\left(\mu^2\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\sum_{i\in[N]}p_i\sum_{l\in[L]}\sum_{j=r_{i,l}+1}^{r_l}\left(\sigma_{l,j}^{(t)}\right)^2\Big] + \zeta^2\right).
\end{aligned}$$

Define $X = \frac{2\Delta_0}{Tq_{\min}} + \frac{2}{q_{\min}}(\chi\epsilon + 2\mu\epsilon^2)$ and $Y = \frac{2C_2}{q_{\min}}$, the function $f(\eta E) = \frac{1}{\eta E}\cdot X + \eta E\cdot Y$ is minimized at $(\eta E)^\star = \sqrt{\frac{X}{Y}}$, and $f((\eta E)^\star) = 2\sqrt{XY}$. Thus, by choosing

$$\eta = \frac{1}{E}\sqrt{\frac{X}{Y}} = \frac{1}{E}\sqrt{\frac{\frac{\Delta_0}{T} + \chi\epsilon + 2\mu\epsilon^2}{C_2}},$$

we balance the first two terms in the bound and have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\left(\mathbb{E}\left[\left\|\nabla_{\mathtt{A}}\mathcal{L}^{\text{lora}}(\mathtt{B}^{(t)},\mathtt{A}^{(t)})\right\|_F^2\right] + \mathbb{E}\left[\left\|\nabla_{\mathtt{B}}\mathcal{L}^{\text{lora}}(\mathtt{B}^{(t)},\mathtt{A}^{(t)})\right\|_F^2\right]\right) \\
&\le 2\sqrt{XY} + \frac{X}{Y}\frac{2C_3}{q_{\min}^2} + \left(\frac{X}{Y}\right)^{\frac{3}{2}}\frac{4C_4}{q_{\min}} + \frac{6C_1}{2q_{\min}^2}\left(\mu^2\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\sum_{i\in[N]}p_i\sum_{l\in[L]}\sum_{j=r_{i,l}+1}^{r_l}\left(\sigma_{l,j}^{(t)}\right)^2\Big] + \zeta^2\right) \\
&= 4\sqrt{\left(\frac{\Delta_0}{Tq_{\min}} + \frac{\chi\epsilon + 2\mu\epsilon^2}{q_{\min}}\right)\cdot\frac{C_2}{q_{\min}}} + \frac{\frac{\Delta_0}{T} + \chi\epsilon + 2\mu\epsilon^2}{C_2}\cdot\frac{2C_3}{q_{\min}^2} + 4\left(\frac{\frac{\Delta_0}{T} + \chi\epsilon + 2\mu\epsilon^2}{C_2}\right)^{\frac{3}{2}}\frac{C_4}{q_{\min}} \\
&\quad + \frac{6C_1}{2q_{\min}^2}\left(\mu^2\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\sum_{i\in[N]}p_i\sum_{l\in[L]}\sum_{j=r_{i,l}+1}^{r_l}\left(\sigma_{l,j}^{(t)}\right)^2\Big] + \zeta^2\right).
\end{aligned}$$

□