Robust Federated Finetuning of LLMs via Alternating Optimization of LoRA

Shuangyi Chen[1*]
University of Toronto
shuangyi.chen@mail.utoronto.ca

Yuanxin Guo[1*]
University of Toronto
yuanxin.guo@mail.utoronto.ca

Yue Ju
Ericsson-GAIA Montréal
yue.ju@ericsson.com

Hardik Dalal
Ericsson-GAIA Montréal
hardik.dalal@ericsson.com

Zhongwen Zhu
Ericsson-GAIA Montréal
zhongwen.zhu@ericsson.com

Ashish Khisti
University of Toronto
akhisti@ece.utoronto.ca

Abstract

Parameter-Efficient Fine-Tuning (PEFT) methods like Low-Rank Adaptation (LoRA) optimize federated training by reducing computational and communication costs. We propose RoLoRA, a federated framework using alternating optimization to fine-tune LoRA adapters. Our approach emphasizes the importance of learning up and down projection matrices to enhance expressiveness and robustness. We use both theoretical analysis and extensive experiments to demonstrate the advantages of RoLoRA over prior approaches that either generate imperfect model updates or limit expressiveness of the model. We provide a theoretical analysis on a linear model to highlight the importance of learning both the down-projection and up-projection matrices in LoRA. We validate the insights on a non-linear model and separately provide a convergence proof under general conditions. To bridge theory and practice, we conducted extensive experimental evaluations on language models including RoBERTa-Large, Llama-2-7B on diverse tasks and FL settings to demonstrate the advantages of RoLoRA over other methods.

1 Introduction

The remarkable performance of large language models (LLMs) stems from their ability to learn at scale. With their broad adaptability and extensive scope, LLMs depend on vast and diverse datasets to effectively generalize across a wide range of tasks and domains. Federated learning [28] offers a promising solution for leveraging data from multiple sources, which could be particularly advantageous for LLMs.

Recently, Parameter-Efficient Fine-Tuning (PEFT) has emerged as an innovative training strategy that updates only a small subset of model parameters, substantially reducing computational and memory demands. A notable method in this category is LoRA [21], which utilizes low-rank matrices to approximate weight changes during fine-tuning. These matrices are integrated with pre-trained weights for inference, facilitating reduced data transfer in scenarios such as federated learning, where update size directly impacts communication efficiency. Many works integrate LoRA into federated

[*]Equal contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

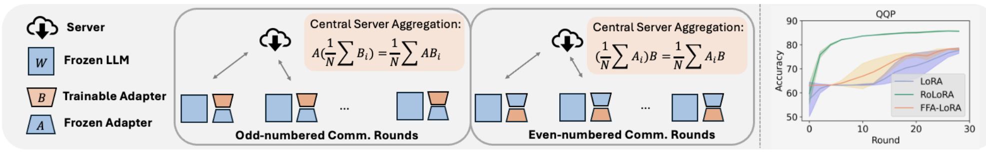

Figure 1: (Left) Overview of the RoLoRA framework. (Right) Performance comparison with baselines on QQP in a 50-client setting, showing RoLoRA's superior convergence speed and final accuracy.

setting [45, 1, 24, 6, 32]. FedPETuning [45] compares various PEFT methods in a federated setting. SLoRA [1] presents a hybrid approach that combines sparse fine-tuning with LoRA to address data heterogeneity in federated settings. Furthermore, FS-LLM [24] presents a framework for fine-tuning LLMs in federated environments. However, these studies typically apply the FedAVG algorithm directly to LoRA modules, resulting in in-exact model updates: the server average of LoRA-A and LoRA-B does not equal the effective adapter.

To address the issue of in-exact model updates, a few recent works have proposed modifications to the down-projection and up-projection components in LoRA. In FlexLoRA [2], the authors propose updating these projections with matrix multiplication followed by truncated SVD. A related method is also considered in FLoRA [40]. Another approach, by Sun et al., introduces a federated finetuning framework named FFA-LoRA [32], which builds on LoRA by freezing the down-projection matrices across all clients and updating only the up-projection matrices. They apply differential privacy [13] to provide privacy guarantees for clients' data. With a sufficient number of finetuning parameters, FFA-LoRA, using a larger learning rate, can achieve performance comparable to FedAVG of LoRA while reducing communication costs by half. However, we observe that with fewer finetuning parameters, FFA-LoRA is less robust than FedAVG of LoRA, primarily due to its reduced expressiveness from freezing down-projections. In this work, we explore the necessity of learning down-projection matrices and propose a federated fine-tuning framework with computational and communication advantages.

We connect the objective of learning down-projection matrices in a federated setting to multitask linear representation learning (MLRL), an approach in which a shared low-rank representation is jointly learned across multiple tasks. While, to the best of our knowledge, the alternating optimization of down- and up-projection matrices has not been explored within the context of LoRA, prior works on MLRL [9, 34] have demonstrated the importance of alternately updating low-rank representations and task-specific heads, demonstrating the necessity of learning a shared representation. Inspired by MLRL, we tackle this challenge by employing alternating optimization for LoRA adapters. We theoretically establish that alternating updates to the two components of LoRA, while maintaining a common global model, enable effective optimization of down-projections and ensure convergence to the global minimizer in a tractable setting.

1.1 Main Contributions

- **RoLoRA framework.** We propose RoLoRA, a robust federated fine-tuning framework based on the alternating optimization of LoRA as shown in Figure 1. RoLoRA fully leverages the expressiveness of LoRA adapters while keeping the computational and communication advantages.
- **Theoretical Insights.** We prove that RoLoRA achieves exponential convergence to the global optimum for federated linear regression, reaching arbitrarily small errors, whereas FFA-LoRA's fixed down-projections result in suboptimality proportional to the initialization error. Simple non-linear model experiments further validate the theoretical importance of updating the down-projections in practice. For the case of smooth non-convex loss functions we also provide a proof of convergence of our proposed method.
- **Empirical results.** Through evaluations on language models (RoBERTa-Large, Llama-2-7B) across various tasks (GLUE, HumanEval, MMLU, Commonsense reasoning tasks), we demonstrate that RoLoRA maintains robustness against reductions in fine-tuning parameters and increases in client numbers compared to prior approaches.

In summary, RoLoRA outperforms baselines with strong theoretical guarantees and robust performance across tasks and scales. It provides new insights into the role of down-projections, advancing the frontier of parameter-efficient federated learning, while its simple yet effective design ensures scalability and communication efficiency.

1.2 Notations

We adopt the notation that lower-case letters represent scalar variables, lower-case bold-face letters denote column vectors, and upper-case bold-face letters denote matrices. The $d \times d$ identity matrix is represented by $\mathbf{I}_d$. Depending on the context, $\|.\|$ denotes the l_2 norm of a vector or the Frobenius norm of a matrix, $\|.\|_{op}$ denotes the operator norm of a matrix, $|.|$ denotes the absolute value of a scalar, $^\top$ denotes matrix or vector transpose. For a number N, $[N] = \{1, \dots, N\}$.

2 Preliminaries and Related Works

2.1 LoRA and Its Variants

Low-Rank Adaptation (LoRA) [21] fine-tunes large language models efficiently by keeping the original weights fixed and adding small trainable matrices that apply low-rank updates. Specifically, for a pre-trained weight matrix $\mathbf{W}_0 \in \mathbb{R}^{d\times d}$, the update is expressed as $\mathbf{W} = \mathbf{W}_0 + \alpha\mathbf{A}\mathbf{B}$, where $\mathbf{A} \in \mathbb{R}^{d\times r}$ and $\mathbf{B} \in \mathbb{R}^{r\times d}$ with $r \ll d$, and only $\mathbf{A}$ and $\mathbf{B}$ are trained. Many variants of LoRA have been developed to further improve efficiency and performance. In Zhang et al. [44], the authors propose a memory-efficient fine-tuning method, LoRA-FA, which keeps the projection-down weight fixed and updates the projection-up weight during fine-tuning. In Zhu et al. [46], the authors highlight the asymmetry between the projection-up and projection-down matrices and focus solely on comparing the effects of freezing either the projection-up or projection-down matrices. Hao et al. [16] introduce the idea of resampling the projection-down matrices, aligning with our observation that freezing projection-down matrices negatively impacts a model's expressiveness. Furthermore, LoRA+ [17] explore the distinct roles of projection-up and projection-down matrices, enhancing performance by assigning different learning rates to each.

2.2 LoRA in Federated Setting

In federated settings, LoRA is practical: clients fine-tune models efficiently with minimal resources, and only the small adapter matrices need to be communicated, sharply reducing transmission costs compared to full model finetuning. Zhang et al. consider FedAVG of LoRA, named FedIT[43]. Zhang et al. [45] compare multiple PEFT methods in the federated setting, including Adapter[20], LoRA[21], Prompt tuning[26] and Bit-Fit[42]. SLoRA[1], which combines sparse finetuning and LoRA, is proposed to address the data heterogeneity in federated setting. As discussed before, Sun et al. [32] design a federated finetuning framework FFA-LoRA by freezing projection-down matrices for all the clients and only updating projection-up matrices. FlexLoRA[2] and FLoRA [40] consider clients with heterogeneous-rank LoRA adapters and proposes federated finetuning approaches. After we completed our work, we noticed a concurrent study, LoRA-A^2 [23], which combines alternating optimization with adaptive rank selection for federated finetuning. While their focus is mainly empirical, the use of alternating optimization in their algorithm is similar to ours.

2.3 FedAVG of LoRA Introduces Interference

Integrating LoRA within a federated setting presents challenges. In such a setup, each of the N clients is provided with the pretrained model weights $\mathbf{W}_0$, which remain fixed during finetuning. Clients are required only to send the updated matrices $\mathbf{B}_i$ and $\mathbf{A}_i$ to a central server for aggregation. While most current studies, such as FedIT[43], SLoRA [1] and FedPETuning [45], commonly apply FedAVG directly to these matrices as shown in (2), this approach might not be optimal. The precise update for each client's model, $\Delta\mathbf{W}_i$, should be calculated as the product of the low-rank matrices $\mathbf{A}_i$ and $\mathbf{B}_i$. Consequently, aggregation on the individual matrices leads to inaccurate model aggregation.

$$\frac{1}{N}\sum_{i=1}^{N}\Delta\mathbf{W}_i = \frac{1}{N}(\mathbf{A}_1\mathbf{B}_1 + \mathbf{A}_2\mathbf{B}_2 + ... + \mathbf{A}_N\mathbf{B}_N) \tag{1}$$

$$\neq \frac{1}{N}(\mathbf{A_1} + \mathbf{A_2} + ... + \mathbf{A_N})\frac{1}{N}(\mathbf{B_1} + \mathbf{B_2} + ... + \mathbf{B_N}) \tag{2}$$

There are a few options to avoid it.

Updating B and A by matrix multiplication and truncated-SVD. One approach [40, 2] involves first computing the product of local matrices $\mathbf{B}_i$ and $\mathbf{A}_i$ to accurately recover $\Delta\mathbf{W}_i$. Then, the global $\mathbf{B}$ and $\mathbf{A}$ of next iteration are obtained by performing truncated SVD on the averaged set of $\Delta\mathbf{W}_i$.

However, this method introduces computational overhead due to the matrix multiplication and SVD operations.

Freezing A (B) during finetuning. Another method is to make clients freeze $\mathbf{B}$ or $\mathbf{A}$ as in Sun et al. [32], leading to precise computation of $\Delta\mathbf{W}$. However, this method limits the expressiveness of the adapter.

With these considerations, we propose a federated finetuning framework, named RoLoRA, based on alternating optimization of LoRA.

3 RoLoRA Framework

In this section, we describe the framework design of RoLoRA and discuss its practical advantages.

Alternating Optimization and Corresponding Aggregation Motivated by the observations discussed in Section 2.3, we propose applying alternating optimization to the local LoRA adapters of each client in a setting with N clients. Unlike the approach in FFA-LoRA, where $\mathbf{A}$ is consistently frozen, we suggest an alternating update strategy. There are alternating odd and even communication rounds designated for updating, aggregating $\mathbf{A}$ and $\mathbf{B}$, respectively.

In the odd-numbered comm. round:

$$\frac{1}{N}\sum_{i=1}^{N}\Delta\mathbf{W}_i^{2t+1}=\frac{1}{N}(\mathbf{A}_1^t\mathbf{B}_1^{t+1}+...+\mathbf{A}_N^t\mathbf{B}_N^{t+1})=\frac{1}{N}\mathbf{A}^t(\mathbf{B}_1^{t+1}+...+\mathbf{B}_N^{t+1}) \tag{3}$$

In the even-numbered comm. round:

$$\frac{1}{N}\sum_{i=1}^{N}\Delta\mathbf{W}_i^{2t+2}=\frac{1}{N}(\mathbf{A}_1^{t+1}\mathbf{B}_1^{t+1}+...+\mathbf{A}_N^{t+1}\mathbf{B}_N^{t+1})=\frac{1}{N}(\mathbf{A}_1^{t+1}+...+\mathbf{A}_N^{t+1})\mathbf{B}^{t+1} \tag{4}$$

As in Algorithm 1 in Appendix, all clients freeze $\mathbf{A}^t$ and update $\mathbf{B}^t$ in the odd communication round. The central server then aggregates these updates to compute $\mathbf{B}^{t+1}=\frac{1}{N}\sum_{i=1}^{N}\mathbf{B}_i^{t+1}$ and distributes $\mathbf{B}^{t+1}$ back to the clients. In the subsequent communication round, clients freeze $\mathbf{B}^{t+1}$ and update $\mathbf{A}^t$. The server aggregates these to obtain $\mathbf{A}^{t+1}=\frac{1}{N}\sum_{i=1}^{N}\mathbf{A}_i^{t+1}$ and returns $\mathbf{A}^{t+1}$ to the clients. It is important to note that in round $2t+1$, the frozen $\mathbf{A}_i^t$ are identical across all clients, as they are synchronized with $\mathbf{A}^t$ from the central server at the beginning of the round. This strategy ensures that the update and aggregation method introduces no interference, as demonstrated in (3) and (4).

Computation and Communication Cost. The parameter-freezing nature of RoLoRA enhances computational and communication efficiency. In each communication round, the number of trainable parameters in the model is effectively halved compared to FedAVG of LoRA. The only additional cost for RoLoRA compared to FFA-LoRA is the alternating freezing of the corresponding parameters. We remark this additional cost is negligible because it is applied to the clients' models and can be executed concurrently during the server's aggregation.

4 Analysis

In this section, we provide an intuitive analysis of why training the down-projection in the LoRA module is essential in a federated setting. We first present a theoretical comparison between RoLoRA and FFA-LoRA on a linear model. While simplified, this comparison offers a direct and rigorous examination of their solutions, clearly highlighting the limitations of FFA-LoRA in this fundamental case. We then empirically validate the theoretical findings using a two-layer non-linear neural network. Finally, we provide a convergence analysis for RoLoRA in non-convex loss landscapes, establishing standard federated learning guarantees.

4.1 Theoretical Insights into Down-Projection Learning: Linear Model Analysis

We start by analyzing RoLoRA and FFA-LoRA within a simplified linear model, offering a clear and rigorous comparison that reveals the inherent limitations of FFA-LoRA's approach. Consider a federated setting with N clients, each with the following local linear model $f_i(\mathbf{X}_i)=\mathbf{X}_i\mathbf{a}\mathbf{b}^\top$ where $\mathbf{Y}_i\in\mathbb{R}^{m\times d}$, $\mathbf{X}_i\in\mathbb{R}^{m\times d}$ with the sample size m, $\mathbf{a}\in\mathbb{R}^d$ (a unit vector) and $\mathbf{b}\in\mathbb{R}^d$ are the LoRA weights corresponding to rank $r = 1$. In this setting, we model the local data of i-th client such that

$$\mathbf{Y}_i = \mathbf{X}_i \mathbf{a}^* \mathbf{b}^{*\top} \tag{5}$$

for some ground truth LoRA weights $\mathbf{a}^* \in \mathbb{R}^d$ (a unit vector) and $\mathbf{b}^* \in \mathbb{R}^d$. We consider the following objective

$$\min_{\mathbf{a} \in \mathbb{R}^d, \mathbf{b} \in \mathbb{R}^d} \frac{1}{N} \sum_{i=1}^{N} l_i(\mathbf{a}, \mathbf{b}) \tag{6}$$

where the local loss is $l_i(\mathbf{a}, \mathbf{b}) = \frac{1}{m} \|\mathbf{X}_i \mathbf{a}^* \mathbf{b}^{*\top} - \mathbf{X}_i \mathbf{a} \mathbf{b}^\top\|^2$. Each $\mathbf{X}_i$ is assumed to be a Gaussian random matrix, where each entry is independently and identically distributed according to a standard Gaussian distribution.

We remind the reader that Section 1.2 provides a summary of mathematical notations and also point to Table 5 in Appendix A3.1 for a summary of the symbols used throughout the theoretical analysis.

Results. In this section, we assume homogeneous clients where there is a single target model as in (5). In the linear model case, we modify RoLoRA from Algorithm 1 to Algorithm 2, employing alternating minimization for $\mathbf{b}$ (line 5) and gradient descent for $\mathbf{a}$ (line 9). Details are described in Algorithm 2 in Appendix. We note that the analysis of the alternating minimization-gradient descent algorithm is inspired by [9, 31, 36] for a different setting of MLRL. See further discussion in Appendix A2.

We aim to show that the training procedure described in Algorithm 2 learns the target model $(\mathbf{a}^*, \mathbf{b}^*)$ by showing the angle distance (Definition 4.2) between $\mathbf{a}$ and $\mathbf{a}^*$ is decreasing in each iteration. Since $\mathbf{b}$ is solved exactly at each iteration via local minimization, it is always optimal with respect to the current $\mathbf{a}$. This allows us to isolate and analyze the convergence behavior of $\mathbf{a}$ using the angle distance, eliminating the potential impact of insufficient local updates of $\mathbf{b}$ on the convergence of $\mathbf{a}$. First, we make typical assumptions on the ground truth $\mathbf{b}^*$ and formally define the angle distance.

Assumption 4.1. There exists $L_{max} < \infty$ (known a priori), s.t. $\|\mathbf{b}^*\| \leq L_{max}$.

Definition 4.2. (Angle Distance) For two unit vectors $\mathbf{a}, \mathbf{a}^* \in \mathbb{R}^d$, the angle distance between $\mathbf{a}$ and $\mathbf{a}^*$ is defined as

$$|\sin \theta(\mathbf{a}, \mathbf{a}^*)| = \|(\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\| \tag{7}$$

where $\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top$ is the projection operator to the direction orthogonal to $\mathbf{a}$.

Let $\delta^t = \|(\mathbf{I}_d - \mathbf{a}^*\mathbf{a}^{*\top})\mathbf{a}^t\| = \|(\mathbf{I}_d - \mathbf{a}^t\mathbf{a}^{t\top})\mathbf{a}^*\|$ denote the angle distance between $\mathbf{a}^*$ and $\mathbf{a}^t$ of t-th iteration. We initialize $\mathbf{a}^0$ such that $|\sin \theta(\mathbf{a}^*, \mathbf{a}^0)| = \delta_0$, where $0 < \delta_0 < 1$, and $\mathbf{b}^0$ is zero. All clients obtain the same initialization for parameters. Now we are ready to state our main results.

Lemma 4.3. *Let $\delta^t = \|(\mathbf{I}_d - \mathbf{a}^*\mathbf{a}^{*\top})\mathbf{a}^t\|$ be the angle distance between $\mathbf{a}^*$ and $\mathbf{a}^t$ of t-th iteration. Assume that Assumption 4.1 holds and $\delta^t \leq \delta^{t-1} \leq \cdots \leq \delta^0$. Let m be the number of samples for each updating step, let auxiliary error thresholds $\epsilon' = \frac{\epsilon_2}{(1-\epsilon_0)(1-\epsilon_1)}$, $\tilde{\epsilon} = \frac{\epsilon_3}{1-\epsilon_0}$ for $\epsilon_0, \epsilon_1, \epsilon_2, \epsilon_3 \in (0, 1)$, if $m = \Omega(q)$ for $q = \max\left(\frac{\log(N)}{[\min(\epsilon_1, \epsilon_2)]^2}, \frac{d \log(\frac{2}{\epsilon_0})}{\epsilon_2^2}\right)$, and auxiliary error thresholds are small such that $\epsilon', \tilde{\epsilon} < \frac{1-(\delta^0)^2}{16}$, for any t and $\eta \leq \frac{1}{L_{max}^2}$, then we have,*

$$\delta^{t+1} \leq \delta^t \sqrt{1 - \eta(1 - \delta^{0^2})\|\mathbf{b}^*\|^2} \tag{8}$$

with probability at least $1 - 2q^{-10}$.

Theorem 4.5 follows by recursively applying Lemma 4.3 and taking a union bound over all $t \in [T]$.

Remark 4.4. The decreasing angle assumption in Lemma 4.3 is a technical tool to simplify the proof. In Theorem 4.5, this condition is not required: the inductive hypothesis inherently enforces the necessary bounds on angles, bypassing the need for explicit monotonicity.

Theorem 4.5. *(Convergence of RoLoRA for linear regressor) Suppose we are in the setting described in Section 4.1 and apply Algorithm 2 for optimization. Given a random initial $\mathbf{a}^0$, an initial angle distance $\delta_0 \in (0, 1)$, we set step size $\eta \leq \frac{1}{L_{max}^2}$ and the number of iterations $T \geq \frac{2}{c(1-(\delta^0)^2)} \log(\frac{\delta^0}{\epsilon})$,* *for $c \in (0,1)$. Under these conditions, if with sufficient number of samples $m = \Omega(q)$ and small auxiliary error thresholds $\epsilon' = \frac{\epsilon_2}{(1-\epsilon_0)(1-\epsilon_1)}, \tilde{\epsilon} = \frac{\epsilon_3}{1-\epsilon_0}$, such that $\epsilon', \tilde{\epsilon} < \frac{1-(\delta^0)^2}{16}$, we achieve that with probability at least $1 - 2Tq^{-10}$ for $q = \max\left(\frac{\log(N)}{[\min(\epsilon_1,\epsilon_2)]^2}, \frac{d\log(\frac{2}{\epsilon_0})}{\epsilon_2^2}\right)$,*

$$\sin\theta(\mathbf{a}^T, \mathbf{a}^*) \le \epsilon$$

which we refer to as ϵ-accurate recovery. In addition,

$$\|\mathbf{a}^T(\mathbf{b}^{T+1})^\top - \mathbf{a}^*(\mathbf{b}^*)^\top\| \le (1+\epsilon')\epsilon\|\mathbf{a}^*\mathbf{b}^{*\top}\|.$$

Theorem 4.5 and Lemma 4.3 show that with a random initialization for the unit vector $\mathbf{a}$ ($\delta^0 \in (0,1)$), RoLoRA makes the global model converge to the target model exponentially fast with large q. The requirement for sample complexity is well-supported, as demonstrated in [10, 12]. While the proof of the above results are relegated to the Appendix, we provide a brief outline of the proof. In Appendices A3.3, we first analyze the minimization step for updating $\mathbf{b}_i^t$ (Lemma A3.9), then establish a bound on the deviation of the gradient from its expectation with respect to $\mathbf{a}$ (Lemma A3.10), and finally derive a bound for $|\sin\theta(\mathbf{a}^{t+1}, \mathbf{a}^*)|$ based on the gradient descent update rule for $\mathbf{a}$ (Lemma 4.3). The proof of Theorem 4.5 is in Section A3.4.

Intuition on Freezing-A Scheme (FFA-LoRA) can Saturate. We begin by applying the FFA-LoRA scheme to a centralized setting, aiming to solve the following optimization problem:

$$\min_{\mathbf{b}\in\mathbb{R}^d} \|\mathbf{X}\mathbf{a}^*\mathbf{b}^{*\top} - \mathbf{X}\mathbf{a}^0\mathbf{b}^\top\|^2$$

where $\mathbf{a}^* \in \mathbb{R}^d$ and $\mathbf{b}^* \in \mathbb{R}^d$ represent the ground truth parameters, and $\mathbf{a}^0 \in \mathbb{R}^d$ is the random initialization. The objective can be transformed to $\sum_{p=1}^d (\mathbf{a}^*b_p^* - \mathbf{a}^0 b_p)^\top \mathbf{X}^\top\mathbf{X}(\mathbf{a}^*b_p^* - \mathbf{a}^0 b_p)$, with b_p as the p-th entry of $\mathbf{b}$, b_p^* as the p-th entry of $\mathbf{b}^*$. In FFA-LoRA scheme, $\mathbf{a}^0$ remains fixed during training. If $\mathbf{a}^0$ is not initialized to be parallel to $\mathbf{a}^*$, the objective can never be reduced to zero. This is because optimizing $\mathbf{b}$ only scales the vector $\mathbf{a}^0 b_p$ along the direction of $\mathbf{a}^0$, without altering the angular distance between $\mathbf{a}^0$ and $\mathbf{a}^*$.

Suppose we are in the federated setting described in Section 4.1, we apply FFA-LoRA, to optimize the objective in (6). In FFA-LoRA scheme, we fix $\mathbf{a}$ of all clients to a random unit vector $\mathbf{a}^0$, where the initial angle distance $\delta^0 = |\sin\theta(\mathbf{a}^*, \mathbf{a}^0)|$, $\delta^0 \in (0,1)$. And we only update $\mathbf{b}_i$ by minimizing l_i and aggregate them.

Proposition 4.6. *(FFA-LoRA lower bound) Suppose we are in the setting described in Section 4.1. For any set of ground truth parameters ($\mathbf{a}^*, \mathbf{b}^*$), the initialization $\mathbf{a}^0$, initial angle distance $\delta^0 \in (0,1)$, we apply FFA-LoRA scheme to obtain a shared global model ($\mathbf{a}^0, \mathbf{b}^{FFA}$), yielding an expected global loss of*

$$\mathbb{E}[\frac{1}{Nm}\sum_{i=1}^N \|\mathbf{X}_i\mathbf{a}^*\mathbf{b}^{*\top} - \mathbf{X}_i\mathbf{a}^0(\mathbf{b}^{FFA})^\top\|^2] = (1+\tilde{c})\|\mathbf{b}^*\|^2(\delta^0)^2 \tag{9}$$

where the expectation is over all the randomness in the $\mathbf{X}_i$, and $\tilde{c} = O(\frac{1}{Nm})$.

See Appendix A3.4.1 for the proof.

Remark 4.7. Proposition 4.6 holds for any unit vector $\mathbf{a}$ and corresponding $\mathbf{b}$ obtained by fully minimizing the local loss. The same expected loss applies to RoLoRA by substituting RoLoRA's reduced angle into Eq. 9.

Comparison of RoLoRA and FFA-LoRA. Proposition 4.6 shows that for any choice of $\delta^0 \in (0,1)$, the global objective reached by FFA-LoRA is shown as in (9). The global objective of FFA-LoRA is dominated by $\|\mathbf{b}^*\|^2(\delta^0)^2$ which is due to the angular distance between $\mathbf{a}^0$ and $\mathbf{a}^*$.

By Theorem 4.5, we demonstrate that RoLoRA achieves ϵ-accurate recovery of the global minimizer. Specifically, the expected global loss of RoLoRA can be upper bounded by $(1+\tilde{c})\|\mathbf{b}^*\|^2\epsilon^2$. Under the same initialization and ground truth parameters for both FFA-LoRA and RoLoRA, RoLoRA's ability to update $\mathbf{a}$ reduces the global loss caused by the angle distance between $\mathbf{a}$ and $\mathbf{a}^*$ from $\|\mathbf{b}^*\|^2(\delta^0)^2$ to $\|\mathbf{b}^*\|^2\epsilon^2$. By increasing the number of iterations, ϵ can be made arbitrarily small.

Heterogeneous Case. In Appendix A3.5, we analyze the convergence of RoLoRA with single LoRA structure in a federated setting with *heterogeneous* clients. By showing the decreasing of the angle distance between the ground truth $\mathbf{a}^*$ and the shared down-projection $\mathbf{a}$, we demonstrate that RoLoRA allows the global model to converge to global minimum while the global loss of FFA-LoRA can be dominated by the term caused by the angle distance between the random initialization $\mathbf{a}^0$ and $\mathbf{a}^*$.

4.2 Verifying Insights On a Non-Linear Model

The previous analysis considers a linear model for each client. To assess the validity of the theorem in a non-linear model, we consider a two-layer neural network model on each client given by

$$f_i(x_i) = \mathsf{ReLU}(x_i\mathbf{A}\mathbf{B})\mathbf{W}_{out} \tag{10}$$

where $\mathbf{W}_{out} \in \mathbb{R}^{d\times c}$, $\mathbf{A} \in \mathbb{R}^{d\times r}$ and $\mathbf{B} \in \mathbb{R}^{r\times d}$ are weights. We train the model on MNIST [11]. We consider two different ways to distribute training images to clients. The first is to distribute the images to 5 clients and each client gets access to training images of two specific labels, while the second is to distribute the images to 10 clients and each client only has training images of one specific label. There is no overlap in the training samples each client can access. Only weights matrices $\mathbf{B}$ and $\mathbf{A}$ are tunable, while $\mathbf{W}_{out}$ are fixed. We use $c = 10, d = 784, r = 16$ and make each client train 5 epochs locally with batch-size 64 and aggregate clients' update following three methods: FedAVG of LoRA, referred as LoRA; FFA-LoRA [32], which freezes $\mathbf{A}$ during training, and RoLoRA, which alternately updates $\mathbf{B}$ and $\mathbf{A}$.

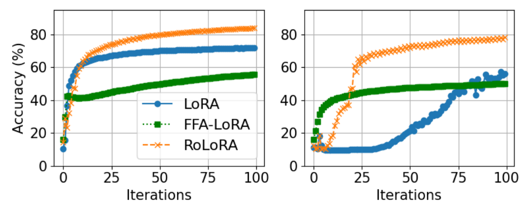

Figure 2: (Left) Comparison of three methods on a toy model with 5 clients. (Right) Comparison of three methods on a toy model with 10 clients.

As shown in Figure 2, we evaluate the performance of the model in each iteration on the test set. We observe that the accuracy of FFA-LoRA plateaus around 55% in both settings, which aligns with our theoretical analysis. The decline in LoRA's performance with an increasing number of clients is most likely due to less accurate model aggregation, as demonstrated in (1) and (2). Notably, RoLoRA demonstrates greater robustness in these settings. *To study the impact of non-linearity on RoLoRA*, we repeat the two-layer experiment on a linear network without ReLU. As shown in Figure 7 in the Appendix, RoLoRA benefits more from the added expressiveness of ReLU.

4.3 Convergence in Smooth Non-Convex Settings

We follow the approach of Li et al.[25] to analyze the convergence behavior of RoLoRA in smooth, non-convex settings, and derive TheoremA4.4 in Appendix. As shown in TheoremA4.4, RoLoRA achieves an $O(1/\sqrt{T})$ convergence rate toward a stationary point under smooth, non-convex conditions, matching the convergence rate established for FedAVG in the same regime.

5 Experiments on Language Models

In this section, we evaluate the performance of RoLoRA in various federated settings. Considering all clients will participate in each round, we will explore the following methods: FedAVG of LoRA (referred as LoRA) [43], FFA-LoRA [32], FlexLoRA[2], FloRA[40], and RoLoRA (ours).

Implementation & Configurations. We implement all the methods based on FederatedScope-LLM [24]. We use NVIDIA GeForce RTX 4090 or NVIDIA A40 for all the experiments. To make a fair comparison, for each dataset, we obtain the best performance on test set and report the average over multiple seeds. Specifically, the learning rate is chosen from the set $\{5e-4, 1e-3, 2e-3, 5e-3, 1e-2, 2e-2, 5e-2, 1e-1\}$. Other hyper-parameters for experiments are specified in Table 6 in Appendix A5.2. Please note that in all tasks, we compare the performance of the three methods *under the same number of communication rounds*.

5.1 Language Understanding Tasks

Model and Datasets. We take the pre-trained RoBERTa-Large (355M) [27] models from the HuggingFace Transformers library, and evaluate the performance of federated finetuning methods on 5 datasets (SST-2, QNLI, MNLI, QQP, RTE) from the GLUE [38].

Rank	Clients Num	Method	SST-2	QNLI	MNLI	QQP	RTE	Avg.
4	3	LoRA	**95.62**±0.17	91.59±0.21	**86.20**±0.05	86.13±0.10	81.46±1.22	88.20
		FFA-LoRA	95.18±0.09	91.35±0.32	84.58±0.21	85.50±0.25	81.10±0.33	87.48
		FlexLoRA	94.91±0.18	90.16±0.49	85.16±0.69	85.69±0.17	79.3±1.05	87.04
		RoLoRA	95.49±0.16	**91.64**±0.30	85.70±0.04	**86.14**±0.06	**82.43**±0.84	**88.28**
4	20	LoRA	94.3±0.27	86.67±2.02	78.55±7.31	83.1±0.04	51.87±3.24	78.90
		FFA-LoRA	93.88±0.06	89.11±0.19	80.99±1.74	83.92±0.2	57.16±1.46	80.01
		FlexLoRA	90.97±1.78	54.36±0.36	53.30±14.59	69.18±10.39	53.19±1.45	64.20
		RoLoRA	**94.88**±0.18	**90.35**±0.37	**85.28**±1.04	**85.83**±0.1	**78.82**±1.7	**87.03**
4	50	LoRA	93.00±0.35	78.13±5.13	52.64±15.07	77.60±1.47	52.23±1.1	70.72
		FFA-LoRA	93.23±0.12	85.05±0.34	69.97±5.57	78.44±0.41	55.72±1.99	76.48
		FlexLoRA	54.08±5.5	55.4±2.03	39.14±2.35	72.00±7.64	52.71±0.00	54.67
		RoLoRA	**94.80**±0.17	**90.00**±0.63	**82.98**±3.36	**85.71**±0.18	**75.57**±2.88	**85.81**
8	50	LoRA	93.00±0.23	79.87±1.52	56.96±2.02	77.45±1.97	53.79±6.57	64.03
		FFA-LoRA	92.74±0.13	83.69±0.75	64.51±1.92	79.71±2.04	53.07±1.3	72.46
		FlexLoRA	50.92±0.00	56.92±1.04	37.43±2.80	66.40±4.74	52.59±0.21	52.85
		RoLoRA	**94.53**±0.17	**90.1**±0.45	**85.17**±0.41	**85.25**±0.13	**76.3**±4.9	**86.27**

Table 1: Results for four methods with RoBERTa-Large models across varying client numbers (3, 20, 50), maintaining a constant sample count during fine-tuning.

Effect of Number of Clients. In Table 1, we increased the number of clients from 3 to 20, and then to 50, ensuring that there is no overlap in the training samples each client can access. Consequently, each client receives a smaller fraction of the total dataset. The configurations are presented in Table 7 in Appendix. We observe that as the number of clients increases, while maintaining the same number of fine-tuning samples, the performance of the LoRA method significantly deteriorates for most datasets. In contrast, RoLoRA maintains its accuracy levels. The performance of FFA-LoRA also declines, attributed to the limited expressiveness of the random initialization of $\mathbf{A}$ for clients' heterogeneous data. FlexLoRA shows significant degradation especially under high client counts. Notably, RoLoRA achieves this accuracy while incurring only half the communication costs associated with LoRA and FlexLoRA. A comparison when using rank-2 LoRA adapter is shown in Table 8 in Appendix.

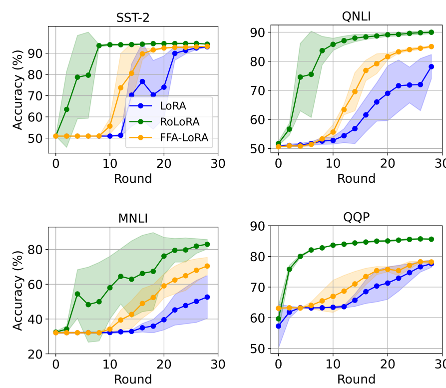

Figure 3: Accuracies over rounds with RoBERTa-Large models on SST-2, QNLI, MNLI, and QQP. It involves 50 clients using rank 4.

Furthermore, we have provided performance comparison of FLoRA and RoLoRA in the same settings in Table 11 in Appendix. RoLoRA consistently outperforms FLoRA across tasks and client counts. Additionally, Figure 3 illustrates the finetuning dynamics, highlighting that RoLoRA converges substantially faster than the other methods. An extended 100-round version is provided in Figure 8 in the Appendix, further demonstrating RoLoRA's superior accuracy when all baselines have converged.

Effect of Non-IID Data Distribution. Table 2 presents a performance comparison of LoRA, FFA-LoRA, FlexLoRA, and RoLoRA on the MNLI and QQP tasks under two federated settings: Dirichlet(0.5) with 10 clients and Dirichlet(1.0) with 15 clients. Here, the Dirichlet(α) parameter controls how non-iid the data is across clients: a smaller α produces highly skewed and heterogeneous client data distributions, while a larger α yields more balanced, moderately non-iid splits. Across both tasks and settings, RoLoRA

	Dir(0.5), #Clients = 10		Dir(1.0), #Clients = 15	
	MNLI	QQP	MNLI	QQP
LoRA	81.19 ±0.23	82.60 ±0.41	74.54 ±1.19	81.49 ±0.60
FFA-LoRA	75.60 ±0.21	81.47 ±0.87	74.83 ±0.59	78.62 ±1.67
FlexLoRA	35.45 ±0.00	63.24 ±0.09	35.45 ±0.00	66.56 ±4.48
RoLoRA	**82.60** ±0.69	**84.16** ±0.65	**81.55** ±0.44	**84.26** ±0.26

Table 2: Performance comparison of methods on MNLI and QQP under different dirichlet distributions and client settings. We report the average and std. over three seeds.

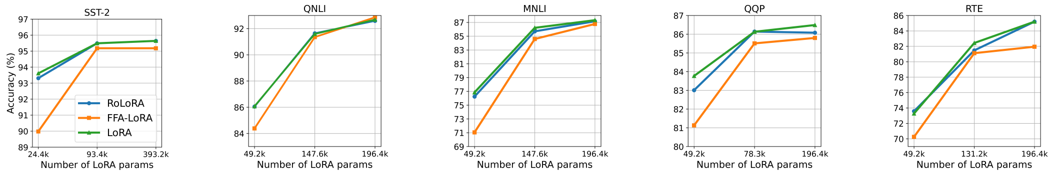

Figure 4: Results with RoBERTa-Large models on GLUE under different fine-tuning parameter budgets, involving three clients with rank 4.

consistently achieves the highest performance. The configurations are presented in Table 7 in Appendix.

Effect of Number of Finetuning Parameters. In Figure 4, we compare three methods across five GLUE datasets. We apply LoRA module to every weight matrix of the selected layers, given different budgets of LoRA parameters. For each dataset, we experiment with three budgets $(\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3)$ ranging from low to high. The corresponding layer sets that are attached with LoRA adapters, $\mathcal{P}_1, \mathcal{P}_2, \mathcal{P}_3$, are detailed in Table 12 in Appendix A5.2. The figures indicates that with sufficient number of finetuning parameters, FFA-LoRA can achieve comparable best accuracy as LoRA and RoLoRA, aligning with the results in [32]; as the number of LoRA parameters is reduced, the performance of the three methods deteriorates to varying degrees. However, RoLoRA, which achieves performance comparable to LoRA, demonstrates greater robustness compared to FFA-LoRA, especially under conditions of limited fine-tuning parameters. It is important to note that with the same finetuning parameters, the communication cost of RoLoRA and FFA-LoRA is always half of that of LoRA due to their parameter freezing nature. This implies that RoLoRA not only sustains its performance but also enhances communication efficiency. Additional results of *varying ranks* are provided in Figure 9, 10, and 11 in Appendix A5.4.

	BoolQ	PIQA	SIQA	HellaSwag
LoRA	61.42$_{\pm 0.29}$	33.19$_{\pm 9.8}$	31.88$_{\pm 3.95}$	21.23$_{\pm 2.82}$
FFA-LoRA	53.43$_{\pm 4.3}$	35.49$_{\pm 9.55}$	10.63$_{\pm 8.44}$	11.81$_{\pm 4.53}$
RoLoRA	**61.83**$_{\pm 0.22}$	**61.26**$_{\pm 3.3}$	**39.76**$_{\pm 0.41}$	**27.49**$_{\pm 2.34}$
	WinoGrande	ARC-e	ARC-c	OBQA
LoRA	31.36$_{\pm 5.02}$	27.36$_{\pm 0.89}$	32.03$_{\pm 1.14}$	26.07$_{\pm 2.32}$
FFA-LoRA	1.61$_{\pm 2.14}$	6.88$_{\pm 0.42}$	7.93$_{\pm 0.89}$	15.0$_{\pm 5.41}$
RoLoRA	**47.67**$_{\pm 0.75}$	**33.19**$_{\pm 1.29}$	**40.13**$_{\pm 1.73}$	**31.67**$_{\pm 1.4}$

Table 3: Results with Llama-2-7B models on commonsense reasoning tasks. This involves 50 clients using rank 8.

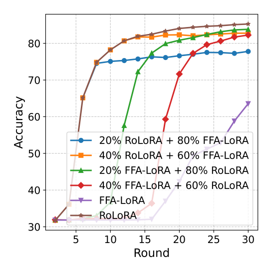

Figure 5: Ablation study on learning vs. freezing A on MNLI task.

5.2 Commonsense Reasoning Tasks

Results. We evaluate RoLoRA against FFA-LoRA and LoRA on Llama-2-7B[35] for commonsense reasoning tasks. In Table 3, we compare the results of three methods with Llama-2-7B models on 8 commonsense reasoning tasks. The configurations are presented in Appendix A5.7. The performance is reported as the mean accuracy with standard deviations across 3 trials. RoLoRA consistently achieves the highest accuracy across all tasks, demonstrating significant improvements over both LoRA and FFA-LoRA. We also highlights that FFA-LoRA exhibits large performance variances across trials, such as a standard deviation of 9.55 for PIQA and 8.44 for SIQA, respectively. This significant variability is likely due to the initialization quality of parameter $\mathbf{A}$, as different initializations could lead to varying optimization trajectories and final performance outcomes as discussed in Section 4. Additional results on this task are presented in Table 16 in Appendix A5.7.

5.3 Ablation Study

Learning vs. Freezing A we conducted an experiment comparing performance of FFA-LoRA, RoLoRA, and different mixing strategies under the setting with 50 clients. In these strategies, for example, 20%RoLoRA+80%FFA-LoRA means we finetune with RoLoRA (where A is learned) for the first 20% of communication rounds, followed by FFA-LoRA (where A is frozen) for the remaining 80%. The results are shown in the Figure 5. We observe that finetuning with RoLoRA generally leads to faster convergence and higher final accuracy. This highlights the benefits of learning A, especially in early training.

Symmetry vs. Asymmetry Update In standard LoRA implementations, LoRA-A is randomly initialized while LoRA-B is set to zero, which implicitly assigns asymmetric roles. In our study, we view LoRA-A as a learnable basis and LoRA-B as coefficients on that basis. Motivated by this, we investigated whether an asymmetric update policy might be preferred. We systematically compare four strategies: (i) the symmetric alternation used in standard RoLoRA, (ii) B-prioritized multi-step updates (B, B, B, A, ...), (iii) A-prioritized multi-step updates (B, A, A, A, ...), and (iv) unequal learning rates (e.g., setting $lr_B = 2lr_A$ or $lr_B = 4lr_A$). As shown in Figure 6, balanced AB alternation yields the highest accuracy and the most stable trajectory, while aggressively prioritizing either A or B degrades performance.

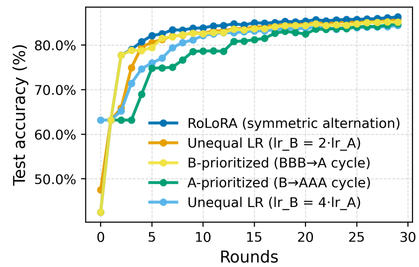

Figure 6: Asymmetry vs. Symmetry in LoRA updates. Accuracy vs. round.

Effect of Local Steps on RoLoRA and FFA-LoRA To evaluate the effect of local steps, we have conducted an ablation study on the number of local steps in a 50-client setting with rank-4 adapters, as shown in Table 4. For a fair comparison, we kept the total computational budget (#Local Steps $\times$ #Total Rounds) constant across all settings. FFA-LoRA's performance consistently degrades on both datasets as the number of local steps increases. This indicates that with more local work per round, FFA-LoRA suffers severely from client drift, where local models overfit to their own data. RoLoRA maintains high performance across all settings.

	(Local Steps, Total Rounds)	(1,600)	(5,120)	(10,60)	(20,30)
MNLI	FFA-LoRA	72.52 $\pm$0.68	71.73 $\pm$1.17	69.64 $\pm$4.31	69.97 $\pm$5.57
	RoLoRA	84.39 $\pm$0.34	84.96 $\pm$0.18	84.79 $\pm$0.23	82.98 $\pm$3.36
QQP	FFA-LoRA	80.51 $\pm$1.38	80.2 $\pm$1.65	79.07 $\pm$1.21	78.44 $\pm$0.41
	RoLoRA	85.24 $\pm$0.56	85.44 $\pm$0.8	84.77 $\pm$0.77	85.71 $\pm$0.18

Table 4: Results on RoBERTa-Large on MNLI and QQP with different local steps while keeping total computational budget constant.

5.4 More Results

We provide additional experimental results in the Appendix, including: (1) evaluations of Llama-2-7B on HumanEval and MMLU tasks (Appendix A5.8); (2) comparisons of communication and time costs in Table 19; and (3) evaluations under a fixed communication budget in Figure 12.

6 Conclusion

In this work, we introduced RoLoRA, a federated finetuning framework based on alternating optimization for LoRA adapters. Our approach addresses key limitations of prior methods by jointly learning both the down-projection and up-projection matrices, thereby enhancing the expressiveness and robustness of the adapted models. Through a combination of theoretical analysis on linear models and validation on nonlinear models, we established the importance of optimizing both components in LoRA. Extensive experimental evaluations across various language models tasks, and diverse federated learning settings confirmed that RoLoRA consistently outperforms existing baselines, particularly in large-scale scenarios under constrained communication and resource conditions.

Acknowledgments and Disclosure of Funding

Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute www.vectorinstitute.ai/partnerships/.

References

[1] Sara Babakniya, Ahmed Elkordy, Yahya Ezzeldin, Qingfeng Liu, Kee-Bong Song, MOSTAFA EL-Khamy, and Salman Avestimehr. SLoRA: Federated parameter efficient fine-tuning of language models. In *International Workshop on Federated Learning in the Age of Foundation Models in Conjunction with NeurIPS 2023*, 2023.

[2] Jiamu Bai, Daoyuan Chen, Bingchen Qian, Liuyi Yao, and Yaliang Li. Federated fine-tuning of large language models under heterogeneous tasks and client resources, 2024.

[3] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for federated learning on user-held data, 2016.

[4] Sahil Chaudhary. Code alpaca: An instruction-following llama model for code generation. `https://github.com/sahil280114/codealpaca`, 2023.

[5] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. 2021.

[6] Shuangyi Chen, Yue Ju, Hardik Dalal, Zhongwen Zhu, and Ashish J Khisti. Robust federated finetuning of foundation models via alternating minimization of loRA. In *Workshop on Efficient Systems for Foundation Models II @ ICML2024*, 2024.

[7] Shuangyi Chen, Yue Ju, Zhongwen Zhu, and Ashish Khisti. Secure inference for vertically partitioned data using multiparty homomorphic encryption. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pages 2508–2513, 2024.

[8] Shuangyi Chen, Anuja Modi, Shweta Agrawal, and Ashish Khisti. Quadratic functional encryption for secure training in vertical federated learning. In *2023 IEEE International Symposium on Information Theory (ISIT)*, pages 60–65, 2023.

[9] Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2089–2099. PMLR, 18–24 Jul 2021.

[10] Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Fedavg with fine tuning: Local updates lead to representation learning, 2022.

[11] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[12] Simon Shaolei Du, Wei Hu, Sham M. Kakade, Jason D. Lee, and Qi Lei. Few-shot learning via learning the representation, provably. In *International Conference on Learning Representations*, 2021.

[13] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer, 2006.

[14] Ahmed El Ouadrhiri and Ahmed Abdelhadi. Differential privacy for deep and federated learning: A survey. *IEEE access*, 10:22359–22380, 2022.

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

[16] Yongchang Hao, Yanshuai Cao, and Lili Mou. Flora: Low-rank adapters are secretly gradient compressors, 2024.

[17] Soufiane Hayou, Nikhil Ghosh, and Bin Yu. Lora+: Efficient low rank adaptation of large models, 2024.

[18] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021.

[19] Erfan Hosseini, Shuangyi Chen, and Ashish Khisti. Secure aggregation in federated learning using multiparty homomorphic encryption. *CoRR*, 2025.

[20] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp, 2019.

[21] 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, 2021.

[22] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, STOC '13, page 665–674, New York, NY, USA, 2013. Association for Computing Machinery.

[23] Jabin Koo, Minwoo Jang, and Jungseul Ok. Towards robust and efficient federated low-rank adaptation with heterogeneous clients, 2024.

[24] 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, 2023.

[25] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*, 2020.

[26] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks, 2022.

[27] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[28] 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*, pages 1273–1282. PMLR, 2017.

[29] Konstantin Mishchenko, Rustem Islamov, Eduard Gorbunov, and Samuel Horváth. Partially personalized federated learning: Breaking the curse of data heterogeneity. *Transactions on Machine Learning Research*, 2025.

[30] Krishna Pillutla, Kshitiz Malik, Abdel-Rahman Mohamed, Mike Rabbat, Maziar Sanjabi, and Lin Xiao. Federated learning with partial model personalization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 17716–17758. PMLR, 17–23 Jul 2022.

[31] Seyedehsara, Nayer, and Namrata Vaswani. Fast and sample-efficient federated low rank matrix recovery from column-wise linear and quadratic projections, 2022.

[32] 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.

[33] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

[34] Kiran Koshy Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Sample efficient linear meta-learning by alternating minimization, 2021.

[35] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[36] Namrata Vaswani. Efficient federated low rank matrix recovery via alternating gd and minimization: A simple proof. *IEEE Transactions on Information Theory*, 70(7):5162–5167, July 2024.

[37] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[38] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding, 2019.

[39] Shuai Wang and Tsung-Hui Chang. Federated matrix factorization: Algorithm design and application to data clustering, 2020.

[40] 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, 2024.

[41] Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H. Yang, Farokhi Farhad, Shi Jin, Tony Q. S. Quek, and H. Vincent Poor. Federated learning with differential privacy: Algorithms and performance analysis, 2019.

[42] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models, 2022.

[43] Jianyi Zhang, Saeed Vahidian, Martin Kuo, Chunyuan Li, Ruiyi Zhang, Tong Yu, Guoyin Wang, and Yiran Chen. Towards building the federatedGPT: Federated instruction tuning. In *International Workshop on Federated Learning in the Age of Foundation Models in Conjunction with NeurIPS 2023*, 2023.

[44] Longteng Zhang, Lin Zhang, Shaohuai Shi, Xiaowen Chu, and Bo Li. Lora-fa: Memory-efficient low-rank adaptation for large language models fine-tuning, 2023.

[45] Zhuo Zhang, Yuanhang Yang, Yong Dai, Qifan Wang, Yue Yu, Lizhen Qu, and Zenglin Xu. FedPETuning: When federated learning meets the parameter-efficient tuning methods of pre-trained language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9963–9977, Toronto, Canada, July 2023. Association for Computational Linguistics.

[46] Jiacheng Zhu, Kristjan Greenewald, Kimia Nadjahi, Haitz Sáez de Ocáriz Borde, Rickard Brüel Gabrielsson, Leshem Choshen, Marzyeh Ghassemi, Mikhail Yurochkin, and Justin Solomon. Asymmetry in low-rank adapters of foundation models, 2024.

Appendix

A1 Algorithms

Algorithm 1 RoLoRA iterations

Input: number of iterations T, number of clients N
for $t = 1$ **to** T **do**
  for $i = 1$ **to** N **do**
    Fix $\mathbf{A}^t$, $\mathbf{B}_i^{t+1} =$ GD-update$(\mathbf{A}^t, \mathbf{B}^t)$
    Transmits $\mathbf{B}_i^{t+1}$ to server
  end for
  Server aggregates $\mathbf{B}^{t+1} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{B}_i^{t+1}$, broadcasts $\mathbf{B}^{t+1}$
  for $i = 1$ **to** N **do**
    Fix $\mathbf{B}^{t+1}$, $\mathbf{A}_i^{t+1} =$ GD-update$(\mathbf{A}^t, \mathbf{B}^{t+1})$
    Transmits $\mathbf{A}_i^{t+1}$ to server
  end for
  Server aggregates $\mathbf{A}^{t+1} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{A}_i^{t+1}$, broadcasts $\mathbf{A}^{t+1}$
end for

Algorithm 2 RoLoRA for linear regressor, Alt-min-GD iterations

Input: GD Step size η, number of iterations T, number of clients N
for $t = 1$ **to** T **do**
  Let $\mathbf{a} \leftarrow \mathbf{a}^{t-1}, \mathbf{b} \leftarrow \mathbf{b}^{t-1}$.
  for $i = 1$ **to** N **do**
    set $\tilde{\mathbf{b}}_i \leftarrow \arg\min_{\mathbf{b}} l_i(\mathbf{a}, \mathbf{b})$
  end for
  $\bar{\mathbf{b}} = \frac{1}{N}\sum_{i=1}^{N} \tilde{\mathbf{b}}_i$
  for $i = 1$ **to** N **do**
    Compute $\nabla_{\mathbf{a}} l_i(\mathbf{a}, \bar{\mathbf{b}})$
  end for
  $\hat{\mathbf{a}}^+ \leftarrow \mathbf{a} - \frac{\eta}{N}\sum_{i=1}^{N} \nabla_{\mathbf{a}} l_i(\mathbf{a}, \bar{\mathbf{b}}),\ \hat{\mathbf{a}} \leftarrow \frac{\hat{\mathbf{a}}^+}{\|\hat{\mathbf{a}}^+\|}$
  $\mathbf{a}^t \leftarrow \hat{\mathbf{a}},\ \mathbf{b}^t \leftarrow \bar{\mathbf{b}}$
end for

A2 Discussion

While Algorithm 2 is conceptually inspired by alternating optimization techniques in matrix sensing and multi-task linear representation learning (MLRL), but it differs in the algorithmic design, and application focus.

Connection to Matrix Sensing. The problem in Eq. (6) is an instance of matrix sensing. Traditional matrix sensing methods [22] focus on recovering low-rank structures from centralized data, whereas RoLoRA is designed for a federated setting, where both data and computation are decentralized across multiple clients. A related line of work is federated matrix factorization, which also applies alternating minimization techniques in distributed environments. Wang et al. [39] perform alternating minimization between local matrix factors within each communication round but do not alternate the aggregation steps. In contrast, RoLoRA alternates both the updates and aggregations of down- and up-projection matrices across rounds, fundamentally changing the communication pattern and mitigating interference between matrix components during aggregation.

Connection to MLRL. As discussed in Section 1, we connect the objective of learning down-projection matrices in a federated setting to multitask linear representation learning (MLRL)[9, 34, 36, 12]. We follow a similar derivation framework to MLRL works that employ alternating optimization, such as FedRep[9] and Vaswani et al. [36]. However, their focus is on a multi-task setting, where the heads are kept diverse and not aggregated. As a result, they perform alternating minimization and gradient descent between the local representation and head within each communication round but do not perform the aggregation steps for the local heads. Furthermore, there are fundamental differences in model structure: in FedRep, the local head acts as a down-projection, whereas in RoLoRA, the corresponding component $\mathbf{B}$ serves as an up-projection. This distinction stems from RoLoRA's foundation in the LoRA framework, where adaptation uses a down-projection followed by an up-projection, with $\mathbf{B}$ mapping back to the original feature space.

Connection to Personalized FL. While RoLoRA is positioned within the global model paradigm, it is worth noting that certain algorithmic structures share interesting similarities. For instance, Mishchenko et al. [29] and Pillutla et al. [30] fall under the category of personalized federated learning, but their update mechanism bears resemblance to RoLoRA's alternating optimization. In [29], clients compute the gradient of their local loss with respect to the global parameters, holding the personalized parameters fixed, and send only this gradient to the server for aggregation. Pillutla et al. [30] adopt an alternating update strategy in the proposed FedAlt algorithm, where clients update personal parameters while holding global parameters fixed, followed by updating the global parameters. Both resemble the alternating structure in RoLoRA. However, RoLoRA operates entirely under a global model setting without personalized components. Furthermore, our alternating scheme is motivated by the decomposition of low-rank adaptation matrices, separating the optimization and also the aggregation of up- and down-projection matrices. Moreover, the underlying proof techniques differ entirely: they employ standard FL convergence analysis, whereas our approach draws on techniques from matrix sensing to highlight the importance of learning the down-projection.

Limitations and Future Works. While RoLoRA demonstrates robust performance across various federated learning settings, our work has a few limitations. The study of learning down-projections has primarily focused on linear models, which may not fully capture the complexities of highly non-linear language models. While empirical validation has been conducted on non-linear models, a rigorous theoretical proof is still lacking. We leave the theoretical extension to simple non-linear models as important future work, although empirical results suggest the method remains effective. Second, the current analysis assumes full client participation in each communication round, which may not hold in real-world federated deployments with intermittent connectivity or resource constraints. A theoretical guarantee for partial client participation is needed.

A3 Theoretical Analysis on Linear Model

A3.1 Notation

Table 5 provides a summary of the symbols used throughout this theoretical analysis.

Notation	Description
$\mathbf{a}^*, \mathbf{b}_i^*$	Ground truth parameters of client i
$\bar{\mathbf{b}}^*$	Average of $\mathbf{b}_i^*$
$\mathbf{a}^t, \mathbf{b}^t$	Global model parameters of t-th iteration
δ_t	The angle distance between $\mathbf{a}^*$ and $\mathbf{a}^t$, $\lvert\sin\theta(\mathbf{a}^*, \mathbf{a}^t)\rvert$
η	Step size
$\mathbf{I_d}$	$d \times d$ identity matrix
$\lVert\cdot\rVert$	l_2 norm of a vector
$\lVert\cdot\rVert_{op}$	Operator norm (l_2 norm) of a matrix
$\lvert\cdot\rvert$	Absolute value of a scalar
$\lVert\cdot\rVert_{\psi_2}$	Sub-Gaussian norm of a sub-Gaussian random variable
$\lVert\cdot\rVert_{\psi_1}$	Sub-exponential norm of a sub-exponential random variable
N	Total number of clients
$\hat{\mathbf{a}}^+$	Updated $\mathbf{a}$ by gradient descent
$\hat{\mathbf{a}}$	Normalized $\hat{\mathbf{a}}^+$
$\tilde{\mathbf{b}}_i$	analytic solution for $\mathbf{b}$ in the local objective function
$\bar{\mathbf{b}}$	Average of $\tilde{\mathbf{b}}_i$

Table 5: Notations

A3.2 Auxiliary

Definition A3.1 (Sub-Gaussian Norm)**.** The sub-Gaussian norm of a random variable ξ, denoted as $\|\xi\|_{\psi_2}$, is defined as:

$$\|\xi\|_{\psi_2} = \inf\{t > 0 : \mathbb{E}[\exp(\xi^2/t^2)] \leq 2\}.$$

A random variable is said to be *sub-Gaussian* if its sub-Gaussian norm is finite. Gaussian random variables are sub-Gaussian. The sub-Gaussian norm of a standard Gaussian random variable $\xi \sim \mathcal{N}(0,1)$ is $\|\xi\|_{\psi_2} = \sqrt{8/3}$.

Definition A3.2 (Sub-Exponential Norm)**.** The sub-exponential norm of a random variable ξ, denoted as $\|\xi\|_{\psi_1}$, is defined as:

$$\|\xi\|_{\psi_1} = \inf\{t > 0 : \mathbb{E}[\exp(|\xi|/t)] \leq 2\}.$$

A random variable is said to be *sub-exponential* if its sub-exponential norm is finite.

Lemma A3.3 (The product of sub-Gaussians is sub-exponential)**.** *Let ξ and υ be sub-Gaussian random variables. Then $\xi\upsilon$ is sub-exponential. Moreover,*

$$\|\xi\upsilon\|_{\psi_1} \leq \|\xi\|_{\psi_2} \cdot \|\upsilon\|_{\psi_2}$$

Lemma A3.4 (Sum of independent sub-Gaussians)**.** *Let $X_1, \cdots, X_N$ be independent mean-zero sub-Gaussian random variables. Then $\sum_{i=1}^N X_i$ is also sub-Gaussian with*

$$\left\|\sum_{i=1}^N X_i\right\|_{\psi_2}^2 \leq C\sum_{i=1}^N \|X_i\|_{\psi_2}^2,$$

where C is some absolute constant.

Proof. See proof of Lemma 2.6.1 of [37]. □

Corollary A3.5. *For random vector $\mathbf{x} \in \mathbb{R}^d$ with entries being independent standard Gaussian random variables, the inner product $\mathbf{a}^\top\mathbf{x}$ is sub-Gaussian for any fixed $\mathbf{a} \in \mathbb{R}^d$, and*

$$\left\|\mathbf{a}^\top\mathbf{x}\right\|_{\psi_2} \leq C'\|\mathbf{a}\|$$

where C' is some absolute constant.

Proof. Note that $\mathbf{a}^\top\mathbf{x} = \sum_{i=1}^d a_i\xi_i$, where $\xi_i \sim \mathcal{N}(0,1)$ is the i-th entry of the random vector $\mathbf{x}$. Choose C to be the absolute constant specified in Lemma A3.4 for standard Gaussian random variables, and set $C' = \sqrt{8C/3}$. We have

$$\left\|\mathbf{a}^\top\mathbf{x}\right\|_{\psi_2}^2 \leq C\sum_{i=1}^N \|a_i\xi_i\|_{\psi_2}^2 \overset{(a)}{=} C\sum_{i=1}^N a_i^2\|\xi_i\|_{\psi_2}^2 \overset{(b)}{=} \frac{8}{3}\cdot C\|\mathbf{a}\|^2 \;\Rightarrow\; \left\|\mathbf{a}^\top\mathbf{x}\right\|_{\psi_2} \leq \sqrt{\frac{8C}{3}}\|\mathbf{a}\| = C'\|\mathbf{a}\|.$$

Step (a) makes use of the homogeneity of the sub-Gaussian norm, and step (b) uses the fact that $\|\xi\|_{\psi_2} = \sqrt{8/3}$ for $\xi \sim \mathcal{N}(0,1)$. □

Definition A3.6 (ϵ-net)**.** Consider a subset $\mathcal{A} \subseteq \mathbb{R}^d$ in the d-dimensional Euclidean space. Let $\epsilon > 0$. A subset $\mathcal{N} \subseteq \mathcal{A}$ is called an ϵ-net of $\mathcal{A}$ if every point of $\mathcal{A}$ is within a distance ϵ of some point in $\mathcal{N}$, i.e.,

$$\forall\, \mathbf{x} \in \mathcal{A},\ \exists\, \mathbf{x}' \in \mathcal{N},\ \|\mathbf{x} - \mathbf{x}'\| \leq \epsilon.$$

Lemma A3.7 (Computing the operator norm on a net)**.** *Let $\mathbf{a} \in \mathbb{R}^d$ and $\epsilon \in [0,1)$. Then, for any ϵ-net $\mathcal{N}$ of the sphere $\mathcal{S}^{d-1}$, we have*

$$\|\mathbf{a}\| \leq \frac{1}{1-\epsilon}\sup_{\mathbf{x}\in\mathcal{N}}\langle\mathbf{a},\mathbf{x}\rangle$$

Proof. See proof of Lemma 4.4.1 of [37]. □

Theorem A3.8 (Bernstein's inequality)**.** *Let $X_1, \ldots, X_N$ be independent mean-zero sub-exponential random variables. Then, for every $t \geq 0$, we have*

$$\mathbb{P}\left(\left|\sum_{i=1}^{N} X_i\right| \geq t\right) \leq 2\exp\left(-c\min\left(\frac{t^2}{\sum_{i=1}^{N}\|X_i\|_{\psi_1}^2}, \frac{t}{\max_i\|X_i\|_{\psi_1}}\right)\right),$$

where $c > 0$ is an absolute constant.

Proof. See proof of Theorem 2.8.1 of [37]. □

A3.3 Homogeneous Case

Outline of Proof. In this section, we first analyze the minimization step for updating $\mathbf{b}_i^t$ (Lemma A3.9), then establish a bound on the deviation of the gradient from its expectation with respect to $\mathbf{a}$ (Lemma A3.10), and finally derive a bound on $|\sin\theta(\mathbf{a}^{t+1}, \mathbf{a}^*)|$ based on the gradient descent update rule for $\mathbf{a}$ (Lemma 4.3 or Lemma A3.11). The proof of Theorem 4.5 is provided in Section A3.4, where the result is obtained by recursively applying Lemma 4.3 over T iterations.

Lemma A3.9. *Let $\mathbf{a} = \mathbf{a}^t$. Let $\delta^t = \|(\mathbf{I}_d - \mathbf{a}^*\mathbf{a}^{*\top})\mathbf{a}\| = \|(\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\|$ denote the angle distance between $\mathbf{a}^*$ and $\mathbf{a}$. Let $\mathbf{g}^\top = \mathbf{a}^\top\mathbf{a}^*\mathbf{b}^{*\top}, \bar{\mathbf{b}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{b}_i$, If $m = \Omega(q)$, and $q = \max(\frac{\log(N)}{[\min(\epsilon_1,\epsilon_2)]^2}, \frac{d\log(\frac{2}{\epsilon_0})}{\epsilon_2^2})$, then with probability $1 - q^{-10}$,*

$$\|\bar{\mathbf{b}} - \mathbf{g}\| \leq \epsilon'\delta^t\|\mathbf{b}^*\| \tag{11}$$

where $\epsilon' = \frac{\epsilon_2}{(1-\epsilon_0)(1-\epsilon_1)}$, for $\epsilon_0, \epsilon_1, \epsilon_2 \in (0, 1)$.

Proof. We drop superscript t for simplicity. Following Algorithm 2, we start by computing the update for $\tilde{\mathbf{b}}_i$. With $\mathbf{g}^\top = \mathbf{a}^\top\mathbf{a}^*\mathbf{b}^{*\top}$, we get:

$$\tilde{\mathbf{b}}_i^\top = \frac{\mathbf{a}^\top\mathbf{X}_i^\top\mathbf{X}_i\mathbf{a}^*\mathbf{b}^{*\top}}{\mathbf{a}^\top\mathbf{X}_i^\top\mathbf{X}_i\mathbf{a}} \tag{12}$$

$$= \frac{\mathbf{a}^\top\mathbf{X}_i^\top\mathbf{X}_i\mathbf{a}\mathbf{a}^\top\mathbf{a}^*\mathbf{b}^{*\top} + \mathbf{a}^\top\mathbf{X}_i^\top\mathbf{X}_i(\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*\top}}{\mathbf{a}^\top\mathbf{X}_i^\top\mathbf{X}_i\mathbf{a}} \tag{13}$$

$$= \mathbf{g}^\top + \frac{\mathbf{a}^\top\mathbf{X}_i^\top\mathbf{X}_i(\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*\top}}{\mathbf{a}^\top\mathbf{X}_i^\top\mathbf{X}_i\mathbf{a}}. \tag{14}$$

Therefore,

$$\|\tilde{\mathbf{b}}_i - \mathbf{g}\| \leq |\mathbf{a}^\top\mathbf{X}_i^\top\mathbf{X}_i\mathbf{a}|^{-1}\cdot\|\mathbf{a}^\top\mathbf{X}_i^\top\mathbf{X}_i(\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*\top}\| = \|\mathbf{X}_i\mathbf{a}\|^{-2}\cdot\|\mathbf{a}^\top\mathbf{X}_i^\top\mathbf{X}_i(\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*\top}\|. \tag{15}$$

Note that since each entry of $\mathbf{X}_i$ is independent and identically distributed according to a standard Gaussian, and $\|\mathbf{a}\| = 1$, $\mathbf{X}_i\mathbf{a}$ is a random standard Gaussian vector. By Theorem 3.1.1 of [37], the following is true for any $\epsilon_1 \in (0, 1)$

$$\mathbb{P}\left\{\|\mathbf{X}_i\mathbf{a}\|^2 \leq (1 - \epsilon_1)m\right\} \leq \exp\left(-\frac{c_1\epsilon_1^2 m}{K^4}\right) \tag{16}$$

where $K = \|\xi\|_{\psi_2} \geq 1$ for $\xi \sim \mathcal{N}(0, 1)$ and c_1 is some large absolute constant that makes (16) holds. Next we upper bound $\|\mathbf{a}^\top\mathbf{X}_i^\top\mathbf{X}_i(\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*\top}\|$. Note that $\mathbb{E}[\mathbf{a}^\top\mathbf{X}_i^\top\mathbf{X}_i(\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*\top}] = \mathbf{a}^\top\mathbb{E}[\mathbf{X}_i^\top\mathbf{X}_i](\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*\top} = m\mathbf{a}^\top(\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*\top} = 0$. First we need to apply sub-exponential Berstein inequality to bound the deviation from this mean, and then apply epsilon net argument. Let $\mathcal{N}$ be any ϵ_0-net of the unit sphere $\mathcal{S}^{d-1}$ in the d-dimensional real Euclidean space, then by Lemma A3.7, we have

$$\|\mathbf{a}^\top\mathbf{X}_i^\top\mathbf{X}_i(\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*\top}\| \leq \frac{1}{1-\epsilon_0}\max_{\mathbf{w}\in\mathcal{N}}\mathbf{a}^\top\mathbf{X}_i^\top\mathbf{X}_i(\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*\top}\mathbf{w} \tag{17}$$

$$\leq \frac{1}{1-\epsilon_0}\max_{\mathbf{w}\in\mathcal{N}}|\mathbf{a}^\top\mathbf{X}_i^\top\mathbf{X}_i(\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*\top}\mathbf{w}| \tag{18}$$

Meanwhile, denote the j-th row of $\mathbf{X}_i$ by $\mathbf{x}_{i,j}^\top$, for every $\mathbf{w} \in \mathcal{N}$, we have

$$\mathbf{a}^\top \mathbf{X}_i^\top \mathbf{X}_i (\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*^\top}\mathbf{w} = \sum_{j=1}^m (\mathbf{a}^\top \mathbf{x}_{i,j})(\mathbf{x}_{i,j}^\top (\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*^\top}\mathbf{w}) \tag{19}$$

On the right hand side of (19), $\mathbf{a}^\top \mathbf{x}_{i,j}$ and $\mathbf{x}_{i,j}^\top (\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*^\top}\mathbf{w}$ are sub-Gaussian random variables. Thus, the summands on the right-hand side of (19) are products of sub-Gaussian random variables, making them sub-exponential. Now by choosing $c_2 = (C')^2$ for the C' in Corollary A3.5, we have the following chain of inequalities for all i, j:

$$\begin{aligned}
\|(\mathbf{a}^\top \mathbf{x}_{i,j})(\mathbf{x}_{i,j}^\top (\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*^\top}\mathbf{w})\|_{\psi_1} &\leq \|\mathbf{a}^\top \mathbf{x}_{i,j}\|_{\psi_2} \cdot \|\mathbf{x}_{i,j}^\top (\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*^\top}\mathbf{w}\|_{\psi_2} && (20)\\
&\leq c_2 \cdot \|\mathbf{a}\| \cdot \|(\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*\top}\mathbf{w}\| && (21)\\
&\leq c_2 \cdot \|\mathbf{a}\| \cdot \|(\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*\top}\|_{op}\|\mathbf{w}\| && (22)\\
&\leq c_2 \cdot \|(\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*\top}\|_{op} && (23)
\end{aligned}$$

Equation (20) is due to Lemma A3.3, (21) is due to Corollary A3.5, (23) is by the fact that $\|\mathbf{a}\| = \|\mathbf{w}\| = 1$.

Furthermore, these summands are mutually independent and have zero mean. By applying sub-exponential Bernstein's inequality (Theorem A3.8) with $t = \epsilon_2 m \|(\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*^\top}\|_{op}$, we get

$$\begin{aligned}
&\mathbb{P}\left\{\left|\mathbf{a}^\top \mathbf{X}_i^\top \mathbf{X}_i (\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*^\top}\mathbf{w}\right| \geq \epsilon_2 m \|(\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*^\top}\|_{op}\right\} && (24)\\
&= \mathbb{P}\left\{\left|\sum_{j=1}^m (\mathbf{a}^\top \mathbf{x}_{i,j})(\mathbf{x}_{i,j}^\top (\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*^\top}\mathbf{w})\right| \geq \epsilon_2 m \|(\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*^\top}\|_{op}\right\} && (25)\\
&= 2\exp\left(-c_3 \epsilon_2^2 m\right) && (26)
\end{aligned}$$

for any fixed $\mathbf{w} \in \mathcal{N}$, $\epsilon_2 \in (0, 1)$, and some absolute constant c_3. (26) follows because

$$\frac{\epsilon_2^2 m^2 \|(\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*^\top}\|_{op}^2}{\sum_{j=1}^m \|(\mathbf{a}^\top \mathbf{x}_{i,j})(\mathbf{x}_{i,j}^\top (\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*\top}\mathbf{w})\|_{\psi_1}^2} \geq \frac{\epsilon_2^2 m}{c_2^2} \tag{27}$$

$$\frac{\epsilon_2 m \|(\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*^\top}\|_{op}}{\max_j \|(\mathbf{a}^\top \mathbf{x}_{i,j})(\mathbf{x}_{i,j}^\top (\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*\top}\mathbf{w})\|_{\psi_1}} \geq \frac{\epsilon_2 m}{c_2} \tag{28}$$

And $\frac{\epsilon_2^2 m}{c_2^2} \leq \frac{\epsilon_2 m}{c_2}$. Now we apply union bound over all elements in $\mathcal{N}$. By Corollary 4.2.13 in [37], there exists an ϵ_0-net $\mathcal{N}$ with $|\mathcal{N}| \leq (\frac{2}{\epsilon_0} + 1)^d$, therefore for this choice of $\mathcal{N}$,

$$\begin{aligned}
&\mathbb{P}\left\{\max_{\mathbf{w}\in\mathcal{N}}\left|\mathbf{a}^\top \mathbf{X}_i^\top \mathbf{X}_i (\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*^\top}\mathbf{w}\right| \geq \epsilon_2 m \|(\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*^\top}\|_{op}\right\} && (29)\\
&\leq \sum_{\mathbf{w}\in\mathcal{N}} \mathbb{P}\left\{\left|\mathbf{a}^\top \mathbf{X}_i^\top \mathbf{X}_i (\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*^\top}\mathbf{w}\right| \geq \epsilon_2 m \|(\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*^\top}\|_{op}\right\} && (30)\\
&\leq \left(\frac{2}{\epsilon_0} + 1\right)^d \cdot 2\exp\left(-c_3 \epsilon_2^2 m\right) && (31)\\
&= 2\exp\left(d\log(1 + \tfrac{2}{\epsilon_0}) - c_3 \epsilon_2^2 m\right) && (32)
\end{aligned}$$

Combining (15),(16), (18), and (32), we get

$$\mathbb{P}\left\{\|\tilde{\mathbf{b}}_i - \mathbf{g}\| \leq \epsilon' \|(\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*^\top}\|_{op}\right\} \geq 1 - p_0 \tag{33}$$

where $\epsilon' = \frac{\epsilon_2}{(1-\epsilon_0)(1-\epsilon_1)}$ and $p_0 = 2\exp\left(d\log(1 + \frac{2}{\epsilon_0}) - c_3 \epsilon_2^2 m\right) + \exp\left(-\frac{c_1 \epsilon_1^2 m}{K^4}\right)$. Using a union bound over $i \in [N]$, we have

$$\mathbb{P}\left\{\bigcap_{i=1}^N \|\tilde{\mathbf{b}}_i - \mathbf{g}\| \leq \epsilon' \|(\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*^\top}\|_{op}\right\} \geq 1 - N p_0. \tag{34}$$

Next we bound $\|\bar{\mathbf{b}} - \mathbf{g}\|$ where $\bar{\mathbf{b}}$ is the average of $\{\mathbf{b}_i\}_{i=1}^N$.

$$\|\bar{\mathbf{b}} - \mathbf{g}\| = \|\frac{1}{N}\sum_{i=1}^{N}(\tilde{\mathbf{b}}_i - \mathbf{g})\| \tag{35}$$

$$\leq \frac{1}{N}\sum_{i=1}^{N}\|\tilde{\mathbf{b}}_i - \mathbf{g}\| \tag{36}$$

$$\leq \epsilon'\|(\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*^\top}\|_{op} \tag{37}$$

$$= \epsilon'\|(\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\| \cdot \|\mathbf{b}^{*^\top}\| \tag{38}$$

$$= \epsilon'\delta^t\|\mathbf{b}^*\| \tag{39}$$

with probability $1 - Np_0$. (36) follows by Jensen's inequality. (38) follows since $\|\mathbf{u}\mathbf{v}^\top\|_{op} = \|\mathbf{u}\| \cdot \|\mathbf{v}\|$. (39) follows since $\delta^t = \|(\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\|$.

If $m = \Omega(q)$, where $q = \max(\frac{\log(N)}{[\min(\epsilon_1,\epsilon_2)]^2}, \frac{d\log(\frac{2}{\epsilon_0})}{\epsilon_2^2})$, then $1 - Np_0 > 1 - \exp(-Cq) > 1 - q^{-10}$ for large absolute constant C. Then with probability at least $1 - q^{-10}$,

$$\|\bar{\mathbf{b}} - \mathbf{g}\| \leq \epsilon'\delta^t\|\mathbf{b}^*\| \tag{40}$$

□

Lemma A3.10. *Let* $\mathbf{a} = \mathbf{a}^t$. *Let* $\delta^t = \|(\mathbf{I}_d - \mathbf{a}^*\mathbf{a}^{*^\top})\mathbf{a}\| = \|(\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\|$ *denote the angle distance between* $\mathbf{a}^*$ *and* $\mathbf{a}$. *Then for* $Nm = \Omega(\frac{d\log(\frac{2}{\epsilon_0})}{\epsilon_3^2})$ *and* $q = \max(\frac{\log(N)}{[\min(\epsilon_1,\epsilon_2)]^2}, \frac{d\log(\frac{2}{\epsilon_0})}{\epsilon_2^2})$, *with probability at least* $1 - 2q^{-10}$,

$$\|\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}}) - \mathbb{E}[\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})]\| \leq 2\tilde{\epsilon}((\epsilon')^2 + \epsilon')\delta^t\|\mathbf{b}^*\|^2 \tag{41}$$

where $\tilde{\epsilon} = \frac{\epsilon_3}{1-\epsilon_0}$, *and* $\epsilon' = \frac{\epsilon_2}{(1-\epsilon_0)(1-\epsilon_1)}$, *for* $\epsilon_0, \epsilon_1, \epsilon_2, \epsilon_3 \in (0, 1)$.

Proof. Based on the loss function $l(\mathbf{a}, \mathbf{b}) = \frac{1}{N}\sum_{i=1}^N l_i(\mathbf{a}, \mathbf{b}) = \frac{1}{Nm}\sum_{i=1}^N\|\mathbf{X}_i\mathbf{a}^*\mathbf{b}^{*^\top} - \mathbf{X}_i\mathbf{a}\mathbf{b}^\top\|^2$, we bound the expected gradient with respect to $\mathbf{a}$ and the deviation from it. The gradient with respect to $\mathbf{a}$ and its expectation are computed as:

$$\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}}) = \frac{2}{Nm}\sum_{i=1}^{N}(\mathbf{X}_i^\top\mathbf{X}_i\mathbf{a}\bar{\mathbf{b}}^\top\bar{\mathbf{b}} - \mathbf{X}_i^\top\mathbf{Y}_i\bar{\mathbf{b}}) \tag{42}$$

$$= \frac{2}{Nm}\sum_{i=1}^{N}(\mathbf{X}_i^\top\mathbf{X}_i\mathbf{a}\bar{\mathbf{b}}^\top\bar{\mathbf{b}} - \mathbf{X}_i^\top\mathbf{X}_i\mathbf{a}^*\mathbf{b}^{*^\top}\bar{\mathbf{b}}) \tag{43}$$

$$= \frac{2}{Nm}\sum_{i=1}^{N}\mathbf{X}_i^\top\mathbf{X}_i(\mathbf{a}\bar{\mathbf{b}}^\top - \mathbf{a}^*\mathbf{b}^{*^\top})\bar{\mathbf{b}} \tag{44}$$

$$\mathbb{E}[\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})] = \frac{2}{Nm}\sum_{i=1}^{N} m(\mathbf{a}\bar{\mathbf{b}}^\top - \mathbf{a}^*\mathbf{b}^{*^\top})\bar{\mathbf{b}} \tag{45}$$

$$= 2(\mathbf{a}\bar{\mathbf{b}}^\top - \mathbf{a}^*\mathbf{b}^{*^\top})\bar{\mathbf{b}} \tag{46}$$

Next, we bound $\|\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}}) - \mathbb{E}[\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})]\|$. Construct ϵ_0-net $\mathcal{N}$ over d dimensional unit spheres $\mathcal{S}^{d-1}$, by Lemma A3.7, we have

$$\|\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}}) - \mathbb{E}[\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})]\| \tag{47}$$

$$\leq \frac{1}{1-\epsilon_0}\max_{\mathbf{w}\in\mathcal{N}}\left|\mathbf{w}^\top\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}}) - \mathbf{w}^\top\mathbb{E}[\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})]\right| \tag{48}$$

$$\leq \frac{1}{1-\epsilon_0}\frac{2}{Nm}\max_{\mathbf{w}\in\mathcal{N}}\left|\sum_{i=1}^{N}\sum_{j=1}^{m}(\mathbf{x}_{i,j}^\top\mathbf{w})(\mathbf{x}_{i,j}(\mathbf{a}\bar{\mathbf{b}}^\top - \mathbf{a}^*\mathbf{b}^{*^\top}))\bar{\mathbf{b}} - \mathbf{w}^\top(\mathbf{a}\bar{\mathbf{b}}^\top - \mathbf{a}^*\mathbf{b}^{*^\top})\bar{\mathbf{b}}\right| \tag{49}$$

where $\mathbf{x}_{i,j}^\top$ is the j-th row of $\mathbf{X}_i$. Observe that $\mathbf{x}_{i,j}^\top \mathbf{w}$ and $\mathbf{x}_{i,j}(\mathbf{a}\bar{\mathbf{b}}^\top - \mathbf{a}^*{\mathbf{b}^*}^\top)\bar{\mathbf{b}}$ are sub-Gaussian variables. Thus the product of them are sub-exponentials. For the right hand side of (49), the summands are sub-exponential random variables with sub-exponential norm

$$
\begin{aligned}
&\|(\mathbf{x}_{i,j}^\top \mathbf{w})(\mathbf{x}_{i,j}(\mathbf{a}\bar{\mathbf{b}}^\top - \mathbf{a}^*{\mathbf{b}^*}^\top))\bar{\mathbf{b}} - \mathbf{w}^\top(\mathbf{a}\bar{\mathbf{b}}^\top - \mathbf{a}^*{\mathbf{b}^*}^\top)\bar{\mathbf{b}}\|_{\psi_1} && (50)\\
&\le \|(\mathbf{x}_{i,j}^\top \mathbf{w})(\mathbf{x}_{i,j}(\mathbf{a}\bar{\mathbf{b}}^\top - \mathbf{a}^*{\mathbf{b}^*}^\top))\bar{\mathbf{b}}\|_{\psi_1} + \|\mathbf{w}^\top(\mathbf{a}\bar{\mathbf{b}}^\top - \mathbf{a}^*{\mathbf{b}^*}^\top)\bar{\mathbf{b}}\|_{\psi_1} && (51)\\
&\le \|(\mathbf{x}_{i,j}^\top \mathbf{w})(\mathbf{x}_{i,j}(\mathbf{a}\bar{\mathbf{b}}^\top - \mathbf{a}^*{\mathbf{b}^*}^\top))\bar{\mathbf{b}}\|_{\psi_1} + \frac{|\mathbf{w}^\top(\mathbf{a}\bar{\mathbf{b}}^\top - \mathbf{a}^*{\mathbf{b}^*}^\top)\bar{\mathbf{b}}|}{\log 2} && (52)\\
&\le c_2 \cdot \|\mathbf{w}\| \cdot \|\mathbf{a}\bar{\mathbf{b}}^\top - \mathbf{a}^*{\mathbf{b}^*}^\top\|_{op} \cdot \|\bar{\mathbf{b}}\| + \frac{|\mathbf{w}^\top(\mathbf{a}\bar{\mathbf{b}}^\top - \mathbf{a}^*{\mathbf{b}^*}^\top)\bar{\mathbf{b}}|}{\log 2} && (53)\\
&\le c_2 \cdot \|\mathbf{w}\| \cdot \|\mathbf{a}\bar{\mathbf{b}}^\top - \mathbf{a}^*{\mathbf{b}^*}^\top\|_{op} \cdot \|\bar{\mathbf{b}}\| + \frac{\|\mathbf{w}\| \cdot \|(\mathbf{a}\bar{\mathbf{b}}^\top - \mathbf{a}^*{\mathbf{b}^*}^\top)\bar{\mathbf{b}}\|}{\log 2} && (54)\\
&\le c_2 \cdot \|\mathbf{w}\| \cdot \|\mathbf{a}\bar{\mathbf{b}}^\top - \mathbf{a}^*{\mathbf{b}^*}^\top\|_{op} \cdot \|\bar{\mathbf{b}}\| + \frac{\|\mathbf{w}\| \cdot \|\mathbf{a}\bar{\mathbf{b}}^\top - \mathbf{a}^*{\mathbf{b}^*}^\top\|_{op} \cdot \|\bar{\mathbf{b}}\|}{\log 2} && (55)\\
&= c_4 \cdot \|\mathbf{a}\bar{\mathbf{b}}^\top - \mathbf{a}^*{\mathbf{b}^*}^\top\|_{op} \cdot \|\bar{\mathbf{b}}\| && (56)
\end{aligned}
$$

where $c_4 = c_2 + \frac{1}{\log 2}$ is some absolute constant greater than 1. Equation (52) is due to the fact that for a constant $c \in \mathbb{R}$,

$$
\|c\|_{\psi_1} = \inf_t \exp\left\{\frac{|c|}{t} \le 2\right\} = \frac{|c|}{\log 2}.
$$

Equation (53) is derived similarly as (20)-(22).

The summands in (49) are mutually independent and have zero mean. Applying sub-exponential Bernstein inequality (Theorem A3.8) with $t = \epsilon_3 Nm \|\mathbf{a}\bar{\mathbf{b}}^\top - \mathbf{a}^*{\mathbf{b}^*}^\top\|_{op} \cdot \|\bar{\mathbf{b}}\|$,

$$
\begin{aligned}
&\mathbb{P}\left\{\left|\sum_{i=1}^N \sum_{j=1}^m [((\mathbf{x}_{i,j}^\top \mathbf{w})(\mathbf{x}_{i,j}(\mathbf{a}\bar{\mathbf{b}}^\top - \mathbf{a}^*{\mathbf{b}^*}^\top)) - \mathbf{w}^\top(\mathbf{a}\bar{\mathbf{b}}^\top - \mathbf{a}^*{\mathbf{b}^*}^\top))\bar{\mathbf{b}}]\right| \ge t\right\} && (57)\\
&\le 2\exp\left(-c\min(\frac{\epsilon_3^2 Nm}{c_4^2}, \frac{\epsilon_3 Nm}{c_4})\right) && (58)\\
&= 2\exp\left(-c_5 \epsilon_3^2 Nm\right) && (59)
\end{aligned}
$$

for any fixed $\mathbf{w} \in \mathcal{N}$, $\epsilon_3 \in (0,1)$ and some absolute constant c_5.

Now we apply union bound over all $\mathbf{w} \in \mathcal{N}$ using Corollary 4.2.13 of [37]. We can conclude that

$$
\begin{aligned}
&\mathbb{P}\left\{\max_{\mathbf{w}\in\mathcal{N}} \left|\mathbf{w}^\top \nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}}) - \mathbf{w}^\top \mathbb{E}[\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})]\right| \ge 2\epsilon_3 \|\mathbf{a}\bar{\mathbf{b}}^\top - \mathbf{a}^*{\mathbf{b}^*}^\top\|_{op} \cdot \|\bar{\mathbf{b}}\|\right\} && (60)\\
&\le \sum_{\mathbf{w}\in\mathcal{N}} \mathbb{P}\left\{\left|\mathbf{w}^\top \nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}}) - \mathbf{w}^\top \mathbb{E}[\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})]\right| \ge 2\epsilon_3 \|\mathbf{a}\bar{\mathbf{b}}^\top - \mathbf{a}^*{\mathbf{b}^*}^\top\|_{op} \cdot \|\bar{\mathbf{b}}\|\right\} && (61)\\
&\le 2\exp\left(d\log(1 + \frac{2}{\epsilon_0}) - c_5 \epsilon_3^2 Nm\right) && (62)
\end{aligned}
$$

Combining (39), (48) , and (60), with probability at least $1-2\exp\left(d\log(1+\frac{2}{\epsilon_0})-c_5\epsilon_3^2 Nm\right)-q^{-10}$,

$$
\begin{aligned}
\|\nabla_{\mathbf{a}} l(\mathbf{a},\bar{\mathbf{b}})-\mathbb{E}[\nabla_{\mathbf{a}} l(\mathbf{a},\bar{\mathbf{b}})]\| &\le \frac{1}{1-\epsilon_0}\max_{\mathbf{w}\in\mathcal{N}}\left|\mathbf{w}^\top\nabla_{\mathbf{a}} l(\mathbf{a},\bar{\mathbf{b}})-\mathbf{w}^\top\mathbb{E}[\nabla_{\mathbf{a}} l(\mathbf{a},\bar{\mathbf{b}})]\right| && (63)\\
&\le \frac{2\epsilon_3}{1-\epsilon_0}\|\mathbf{a}\bar{\mathbf{b}}^\top-\mathbf{a}^*\mathbf{b}^{*\top}\|_{op}\cdot\|\bar{\mathbf{b}}\| && (64)\\
&= \frac{2\epsilon_3}{1-\epsilon_0}\|\mathbf{a}(\bar{\mathbf{b}}-\mathbf{g})^\top-(\mathbf{I}_d-\mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*\top}\|_{op}\cdot\|\bar{\mathbf{b}}\| && (65)\\
&\le \frac{2\epsilon_3}{1-\epsilon_0}(\|\mathbf{a}^\top(\bar{\mathbf{b}}-\mathbf{g})\|+\|(\mathbf{I}_d-\mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*\top}\|_{op})\|\bar{\mathbf{b}}\| && (66)\\
&\le \frac{2\epsilon_3}{1-\epsilon_0}(\|\bar{\mathbf{b}}-\mathbf{g}\|+\|(\mathbf{I}_d-\mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\|\cdot\|\mathbf{b}^*\|)\|\bar{\mathbf{b}}\| && (67)\\
&= \frac{2\epsilon_3}{1-\epsilon_0}(\|\bar{\mathbf{b}}-\mathbf{g}\|+\delta^t\|\mathbf{b}^*\|)\|\bar{\mathbf{b}}-\mathbf{g}+\mathbf{g}\| && (68)\\
&\le \frac{2\epsilon_3}{1-\epsilon_0}(\|\bar{\mathbf{b}}-\mathbf{g}\|+\delta^t\|\mathbf{b}^*\|)(\|\bar{\mathbf{b}}-\mathbf{g}\|+\|\mathbf{g}\|) && (69)\\
&\le \frac{2\epsilon_3}{1-\epsilon_0}(\|\bar{\mathbf{b}}-\mathbf{g}\|+\delta^t\|\mathbf{b}^*\|)(\|\bar{\mathbf{b}}-\mathbf{g}\|+\|\mathbf{b}^*\|) && (70)\\
&= \frac{2\epsilon_3}{1-\epsilon_0}(\|\bar{\mathbf{b}}-\mathbf{g}\|^2+\delta^t\|\bar{\mathbf{b}}-\mathbf{g}\|\|\mathbf{b}^*\|+\|\bar{\mathbf{b}}-\mathbf{g}\|\|\mathbf{b}^*\|+\delta^t\|\mathbf{b}^*\|^2) && (71)\\
&\le \frac{2\epsilon_3}{1-\epsilon_0}((\epsilon')^2(\delta^t)^2+\epsilon'(\delta^t)^2+\epsilon'\delta^t+\delta^t)\|\mathbf{b}^*\|^2 && (72)\\
&\le \frac{2\epsilon_3}{1-\epsilon_0}(\epsilon'+1)^2\delta^t\|\mathbf{b}^*\|^2 && (73)\\
&= 2\tilde{\epsilon}(\epsilon'+1)^2\delta^t\|\mathbf{b}^*\|^2 && (74)
\end{aligned}
$$

with $\tilde{\epsilon}=\frac{\epsilon_3}{1-\epsilon_0}$. (64) uses (60). (66) follows by triangle inequality. (68) follows by $\delta^t=\|(\mathbf{I}_d-\mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\|$. (70) uses $\|\mathbf{g}\|=\|\mathbf{b}^*\mathbf{a}^{*\top}\mathbf{a}\|\le\|\mathbf{b}^*\|$. (73) follows by $(\delta^t)^2<\delta^t$ since $\delta^t\in(0,1)$.

If $Nm=\Omega(\frac{d\log(\frac{2}{\epsilon_0})}{\epsilon_3^2})$, then existing large constant C,

$$
\begin{aligned}
1-2\exp\left(d\log(1+\frac{2}{\epsilon_0})-c_5\epsilon_3^2 Nm\right)-q^{-10} &> 1-\exp(-Cd)-q^{-10} && (75)\\
&> 1-d^{-10}-q^{-10} && (76)\\
&> 1-2q^{-10} && (77)
\end{aligned}
$$

Thus with probability at least $1-2q^{-10}$, (74) holds. □

Lemma A3.11 (Lemma 4.3)**.** *Let $\mathbf{a}=\mathbf{a}^t$. Let $\delta^t=\|(\mathbf{I}_d-\mathbf{a}^*\mathbf{a}^{*\top})\mathbf{a}\|=\|(\mathbf{I}_d-\mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\|$ denote the angle distance between $\mathbf{a}^*$ and $\mathbf{a}$. Assume that Assumption 4.1 holds and $\delta^t\le\delta^{t-1}\le\cdots\le\delta^0$. Let m be the number of samples for each updating step, let $\epsilon'=\frac{\epsilon_2}{(1-\epsilon_0)(1-\epsilon_1)}$, $\tilde{\epsilon}=\frac{\epsilon_3}{1-\epsilon_0}$ for $\epsilon_0,\epsilon_1,\epsilon_2,\epsilon_3\in(0,1)$, if*

$$
m=\Omega\left(\max\left\{\frac{\log(N)}{[\min(\epsilon_1,\epsilon_2)]^2},\frac{d\log(\frac{2}{\epsilon_0})}{\epsilon_2^2}\right\}\right)
$$

and $\epsilon',\tilde{\epsilon}<\frac{1-(\delta^0)^2}{16}$, for any t and $\eta\le\frac{1}{L_{max}^2}$, then we have,

$$
\delta^{t+1}\le\delta^t\sqrt{1-\eta(1-\delta^{02})\|\mathbf{b}^*\|^2} \tag{78}
$$

with probability at least $1-2q^{-10}$ for $q=\max\left\{\frac{\log(N)}{[\min(\epsilon_1,\epsilon_2)]^2},\frac{d\log(\frac{2}{\epsilon_0})}{\epsilon_2^2}\right\}$.

Proof. Recall that $\hat{\mathbf{a}}^+ = \mathbf{a} - \eta\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})$. We substract and add $\mathbb{E}[\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})]$, obtain

$$\hat{\mathbf{a}}^+ = \mathbf{a} - \eta\mathbb{E}[\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})] + \eta(\mathbb{E}[\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})] - \nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})) \tag{79}$$

Multiply both sides by the projection operator $\mathbf{P} = \mathbf{I}_d - \mathbf{a}^*(\mathbf{a}^*)^\top$,

$$\begin{aligned}
\mathbf{P}\hat{\mathbf{a}}^+ &= \mathbf{P}\mathbf{a} - \eta\mathbf{P}\mathbb{E}[\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})] + \eta\mathbf{P}(\mathbb{E}[\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})] - \nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})) &(80)\\
&= \mathbf{P}\mathbf{a} - 2\eta\mathbf{P}(\mathbf{a}\bar{\mathbf{b}}^\top - \mathbf{a}^*\mathbf{b}^{*^\top})\bar{\mathbf{b}} + \eta\mathbf{P}(\mathbb{E}[\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})] - \nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})) &(81)\\
&= \mathbf{P}\mathbf{a} - 2\eta\mathbf{P}\mathbf{a}\bar{\mathbf{b}}^\top\bar{\mathbf{b}} + \eta\mathbf{P}(\mathbb{E}[\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})] - \nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})) &(82)\\
&= \mathbf{P}\mathbf{a}(1 - 2\eta\bar{\mathbf{b}}^\top\bar{\mathbf{b}}) + \eta\mathbf{P}(\mathbb{E}[\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})] - \nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})) &(83)
\end{aligned}$$

where (81) uses $\mathbb{E}[\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})] = 2(\mathbf{a}\bar{\mathbf{b}}^\top - \mathbf{a}^*\mathbf{b}^{*^\top})\bar{\mathbf{b}}$, (82) follows by $\mathbf{P}\mathbf{a}^* = 0$. Thus, we get

$$\|\mathbf{P}\hat{\mathbf{a}}^+\| \leq \|\mathbf{P}\mathbf{a}\||1 - 2\eta\bar{\mathbf{b}}^\top\bar{\mathbf{b}}| + \eta\|(\mathbb{E}[\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})] - \nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}}))\| \tag{84}$$

Normalizing the left hand side, we obtain

$$\begin{aligned}
\frac{\|\mathbf{P}\hat{\mathbf{a}}^+\|}{\|\hat{\mathbf{a}}^+\|} &\leq \frac{\|\mathbf{P}\mathbf{a}\||1 - 2\eta\bar{\mathbf{b}}^\top\bar{\mathbf{b}}| + \eta\|(\mathbb{E}[\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})] - \nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}}))\|}{\|\hat{\mathbf{a}}^+\|} &(85)\\
\Rightarrow \delta^{t+1} &\leq \frac{\delta^t|1 - 2\eta\bar{\mathbf{b}}^\top\bar{\mathbf{b}}| + \eta\|\mathbb{E}[\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})] - \nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})\|}{\|\hat{\mathbf{a}}^+\|} &(86)\\
&= \frac{E_1 + E_2}{\|\hat{\mathbf{a}}^+\|} &(87)
\end{aligned}$$

where (86) follows by $\delta^{t+1} = \frac{\|\mathbf{P}\hat{\mathbf{a}}^+\|}{\|\hat{\mathbf{a}}^+\|}$ and $\delta^t = \|\mathbf{P}\mathbf{a}\|$. We need to upper bound E_1 and E_2 accordingly. E_2 is upper bounded based on Lemma A3.10. With probability at least $1 - 2q^{-10}$,

$$\begin{aligned}
E_2 &= \eta\|\mathbb{E}[\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})] - \nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})\| &(88)\\
&\leq 2\eta\tilde{\epsilon}(\epsilon' + 1)^2\delta^t\|\mathbf{b}^*\|^2 &(89)
\end{aligned}$$

To upper bound E_1, we need to lower bound $\|\bar{\mathbf{b}}\|^2$. We can first lower bound $\|\bar{\mathbf{b}}\|$ by:

$$\begin{aligned}
\|\bar{\mathbf{b}}\| &= \|\mathbf{g} - (\mathbf{g} - \bar{\mathbf{b}})\| &(90)\\
&\geq \|\mathbf{g}\| - \|\mathbf{g} - \bar{\mathbf{b}}\| &(91)\\
&= \sqrt{1 - (\delta^t)^2}\|\mathbf{b}^*\| - \|\mathbf{g} - \bar{\mathbf{b}}\| &(92)\\
&\geq \sqrt{1 - (\delta^t)^2}\|\mathbf{b}^*\| - \epsilon'\delta^t\|\mathbf{b}^*\| &(93)
\end{aligned}$$

with probability at least $1 - q^{-10}$. (92) follows by $\mathbf{g}^\top = \mathbf{a}^\top\mathbf{a}^*\mathbf{b}^{*^\top}$ and $\mathbf{a}^\top\mathbf{a}^* = \cos\theta(\mathbf{a}, \mathbf{a}^*)$, (93) follows by Lemma A3.9. Assuming $\delta^t \leq \cdots \leq \delta^0$, we choose $\epsilon' < \frac{1-(\delta^0)^2}{16}$ to make $\sqrt{1 - (\delta^t)^2}\|\mathbf{b}^*\| - \epsilon'\delta^t\|\mathbf{b}^*\| \geq 0$. Hence $\|\bar{\mathbf{b}}\|^2$ is lower bounded by:

$$\begin{aligned}
\|\bar{\mathbf{b}}\|^2 &\geq (\sqrt{1 - (\delta^t)^2}\|\mathbf{b}^*\| - \epsilon'\delta^t\|\mathbf{b}^*\|)^2 &(94)\\
&= (1 - (\delta^t)^2)\|\mathbf{b}^*\|^2 + (\epsilon')^2(\delta^t)^2\|\mathbf{b}^*\|^2 - 2\epsilon'\delta^t\sqrt{1 - (\delta^t)^2}\|\mathbf{b}^*\|^2 &(95)\\
&\geq (1 - (\delta^t)^2)\|\mathbf{b}^*\|^2 + (\epsilon')^2(\delta^t)^2\|\mathbf{b}^*\|^2 - \epsilon'\|\mathbf{b}^*\|^2 &(96)\\
&\geq (1 - (\delta^0)^2)\|\mathbf{b}^*\|^2 - \epsilon'\|\mathbf{b}^*\|^2 &(97)
\end{aligned}$$

with probability at least $1 - q^{-10}$. (96) follows by $xy \leq \frac{1}{2}$ for $x^2 + y^2 = 1$, (97) follows by assuming $\delta^t \leq \delta^{t-1} \leq \cdots \leq \delta^0$. E_1 is upper bounded below.

$$\begin{aligned}
E_1 &= \delta^t|1 - 2\eta\bar{\mathbf{b}}^\top\bar{\mathbf{b}}| &(98)\\
&\leq \delta^t|1 - 2\eta((1 - (\delta^0)^2) - \epsilon')\|\mathbf{b}^*\|^2| &(99)
\end{aligned}$$

with probability at least $1 - q^{-10}$. Next we lower bound $\|\hat{\mathbf{a}}^+\|$.

$$\begin{aligned}
\|\hat{\mathbf{a}}^+\|^2 &= \|\mathbf{a} - \eta\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})\|^2 &(100)\\
&= \mathbf{a}^\top\mathbf{a} + \eta^2\|\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})\|^2 - 2\eta\mathbf{a}^\top\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}}) &(101)\\
&\geq \mathbf{a}^\top\mathbf{a} - 2\eta\mathbf{a}^\top\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}}) &(102)\\
&= 1 - 2\eta\mathbf{a}^\top\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}}) &(103)\\
&= 1 - 2\eta\mathbf{a}^\top(\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}}) - \mathbb{E}[\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})]) - 2\eta\mathbf{a}^\top\mathbb{E}[\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})] &(104)
\end{aligned}$$

where (102) follows by $\eta^2\|\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})\|^2 \geq 0$, and (103) follows by $\mathbf{a}^\top\mathbf{a} = 1$. The first subtrahend $2\eta\mathbf{a}^\top(\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}}) - \mathbb{E}[\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})])$ is upper bounded such that

$$
\begin{aligned}
2\eta\mathbf{a}^\top(\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}}) - \mathbb{E}[\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})]) &\leq 2\eta\|\mathbf{a}\| \cdot \|(\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}}) - \mathbb{E}[\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})])\| && (105)\\
&= 2\eta\|(\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}}) - \mathbb{E}[\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})])\| && (106)\\
&\leq 4\eta\tilde{\epsilon}(\epsilon' + 1)^2\|\mathbf{b}^*\|^2 && (107)
\end{aligned}
$$

with probability at least $1 - 2q^{-10}$. (107) uses Lemma A3.9. The second subtrahend is upper bounded such that

$$
\begin{aligned}
2\eta\mathbf{a}^\top\mathbb{E}[\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})] &= 4\eta\mathbf{a}^\top(\mathbf{a}\bar{\mathbf{b}}^\top - \mathbf{a}^*\mathbf{b}^{*\top})\bar{\mathbf{b}} && (108)\\
&= 4\eta\mathbf{a}^\top(\mathbf{a}\bar{\mathbf{b}}^\top - \mathbf{a}^*\mathbf{b}^{*\top})\mathbf{g} - 4\eta\mathbf{a}^\top(\mathbf{a}\bar{\mathbf{b}}^\top - \mathbf{a}^*\mathbf{b}^{*\top})(\mathbf{g} - \bar{\mathbf{b}}) && (109)
\end{aligned}
$$

where $\mathbf{a}^\top(\mathbf{a}\bar{\mathbf{b}}^\top - \mathbf{a}^*\mathbf{b}^{*\top})\mathbf{g} = -\mathbf{a}^\top((\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*\top} + \mathbf{a}(\mathbf{g} - \bar{\mathbf{b}})^\top)\mathbf{g} = (\bar{\mathbf{b}} - \mathbf{g})^\top\mathbf{g}$. The second term is simplified via $\mathbf{a}^\top(\mathbf{a}\bar{\mathbf{b}}^\top - \mathbf{a}^*\mathbf{b}^{*\top})(\mathbf{g} - \bar{\mathbf{b}}) = \mathbf{a}^\top((\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top)\mathbf{a}^*\mathbf{b}^{*\top} + \mathbf{a}(\mathbf{g} - \bar{\mathbf{b}})^\top)(\bar{\mathbf{b}} - \mathbf{g}) = -(\mathbf{g} - \bar{\mathbf{b}})^2$. Both simplifications use $\mathbf{a}^\top(\mathbf{I}_d - \mathbf{a}\mathbf{a}^\top) = 0$ and $\mathbf{a}^\top\mathbf{a} = 1$. (109) becomes

$$
\begin{aligned}
2\eta\mathbf{a}^\top\mathbb{E}[\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})] &= 4\eta(\bar{\mathbf{b}} - \mathbf{g})^\top\mathbf{g} + 4\eta(\mathbf{g} - \bar{\mathbf{b}})^2 && (110)\\
&\leq 4\eta\|\mathbf{g} - \bar{\mathbf{b}}\|\|\mathbf{b}^*\| + 4\eta\|\mathbf{g} - \bar{\mathbf{b}}\|^2 && (111)\\
&\leq 4\eta\epsilon'\delta^t\|\mathbf{b}^*\|^2 + 4\eta(\epsilon')^2(\delta^t)^2\|\mathbf{b}^*\|^2 && (112)\\
&\leq 4\eta((\epsilon')^2 + \epsilon')\|\mathbf{b}^*\|^2 && (113)
\end{aligned}
$$

with probability at least $1 - q^{-10}$. (112) uses Lemma A3.9. Combining (107) and (113), we obtain

$$
\|\hat{\mathbf{a}}^+\|^2 \geq 1 - 4\eta\tilde{\epsilon}(\epsilon' + 1)^2\|\mathbf{b}^*\|^2 - 4\eta((\epsilon')^2 + \epsilon')\|\mathbf{b}^*\|^2 \quad (114)
$$

with probability at least $1 - 2q^{-10}$. Combining (99), (89) and (114), we obtain

$$
\delta^{t+1} \leq \frac{E_1 + E_2}{\|\hat{\mathbf{a}}^+\|} \leq \frac{\delta^t(1 - 2\eta((1 - (\delta^0)^2) - \epsilon')\|\mathbf{b}^*\|^2 + 2\eta\tilde{\epsilon}(\epsilon' + 1)^2\|\mathbf{b}^*\|^2)}{\sqrt{1 - 4\eta\tilde{\epsilon}(\epsilon' + 1)^2\|\mathbf{b}^*\|^2 - 4\eta((\epsilon')^2 + \epsilon')\|\mathbf{b}^*\|^2}} = \delta^t C \quad (115)
$$

We can choose $\epsilon', \tilde{\epsilon} < \frac{1-(\delta^0)^2}{16}$ such that $(1-(\delta^0)^2) > \max(4(\tilde{\epsilon}(\epsilon'+1)^2+(\epsilon')^2+\epsilon'), 2\epsilon'+2\tilde{\epsilon}(\epsilon'+1)^2)$ holds. Then we obtain

$$
\begin{aligned}
C &= \frac{1 - 2\eta((1 - (\delta^0)^2) - \epsilon')\|\mathbf{b}^*\|^2 + 2\eta\tilde{\epsilon}(\epsilon' + 1)^2\|\mathbf{b}^*\|^2}{\sqrt{1 - 4\eta\tilde{\epsilon}(\epsilon' + 1)^2\|\mathbf{b}^*\|^2 - 4\eta((\epsilon')^2 + \epsilon')\|\mathbf{b}^*\|^2}} && (116)\\
&= \frac{1 - 2\eta(1 - (\delta^0)^2)\|\mathbf{b}^*\|^2 + 2\eta\epsilon'\|\mathbf{b}^*\|^2 + 2\eta\tilde{\epsilon}(\epsilon' + 1)^2\|\mathbf{b}^*\|^2}{\sqrt{1 - 4\eta(\tilde{\epsilon}(\epsilon' + 1)^2 + (\epsilon')^2 + \epsilon')\|\mathbf{b}^*\|^2}} && (117)\\
&\leq \frac{1 - 2\eta(1 - (\delta^0)^2)\|\mathbf{b}^*\|^2 + \eta(2\epsilon' + 2\tilde{\epsilon}(\epsilon' + 1)^2)\|\mathbf{b}^*\|^2}{\sqrt{1 - 4\eta(\tilde{\epsilon}(\epsilon' + 1)^2 + (\epsilon')^2 + \epsilon')\|\mathbf{b}^*\|^2}} && (118)\\
&\leq \frac{1 - \eta(1 - (\delta^0)^2)\|\mathbf{b}^*\|^2}{\sqrt{1 - \eta(1 - (\delta^0)^2)\|\mathbf{b}^*\|^2}} && (119)\\
&= \sqrt{1 - \eta(1 - (\delta^0)^2)\|\mathbf{b}^*\|^2} && (120)
\end{aligned}
$$

Assuming $\eta \leq \frac{1}{L_{max}^2} \leq \frac{1}{\|\mathbf{b}^*\|^2}$, $1 - \eta(1 - (\delta^0)^2)\|\mathbf{b}^*\|^2$ is strictly positive. Therefore we obtain, with probability at least $1 - 2q^{-10}$,

$$
\delta^{t+1} \leq \delta^t\sqrt{1 - \eta(1 - (\delta^0)^2)\|\mathbf{b}^*\|^2}. \quad (121)
$$

□

A3.4 Proof of Theorem 5.4

Proof. In Lemma 4.3, we have shown the angle distance between $\mathbf{a}$ and $\mathbf{a}^*$ decreasing in t-th iteration such that with probability at least $1 - 2q^{-10}$ for $q = \max\{\log(N), d\}$, $\delta^{t+1} \leq \delta^t C$ for $c \in (0, 1), C = \sqrt{1 - c(1 - (\delta^0)^2)}$ with proper choice of step size η.

Proving $\delta^1 \leq \delta^0 C$. Now we are to prove that for the first iteration, $\delta^1 \leq \delta^0 C$ with certain probability.

By Lemma A3.9, we get $\|\bar{\mathbf{b}} - \mathbf{g}\| \leq \epsilon' \delta^0 \|\mathbf{b}^*\|$ with probability at least $1 - q^{-10}$.

By Lemma A3.10, we get $\|\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}}) - \mathbb{E}[\nabla_{\mathbf{a}} l(\mathbf{a}, \bar{\mathbf{b}})]\| \leq 2\tilde{\epsilon}((\epsilon')^2 + \epsilon')\delta^0 \|\mathbf{b}^*\|^2$ with probability at least $1 - 2q^{-10}$.

By Lemma A3.11, without assuming decreasing angles, we obtain, with probability at least $1 - 2q^{-10}$,

$$\delta^1 \leq \delta^0 \sqrt{1 - \eta(1 - (\delta^0)^2)\|\mathbf{b}^*\|^2}. \tag{122}$$

Inductive Hypothesis. Based on inductive hypothesis, by proving $\delta^1 \leq \delta^0 C$, the assumption that $\delta^t \leq \delta^{t-1} C \leq \cdots \leq \delta^1 C^{t-1}$, and proving $\delta^{t+1} \leq \delta^t C$, we conclude that $\delta^t \leq \delta^{t-1} C$ holds for all $t \in [T]$ iterations. We take a union bound over all $t \in [T]$ such that,

$$\mathbb{P}\left\{\bigcap_{t=0}^{T-1} \delta^{t+1} \leq \delta^t \sqrt{1 - c(1 - (\delta^0)^2)}\right\} \geq 1 - 2Tq^{-10}. \tag{123}$$

Solve for T. In order to achieve ϵ-recovery of $\mathbf{a}^*$, we need

$$\delta^0 (1 - c(1 - (\delta^0)^2))^{\frac{T}{2}} \leq \epsilon \tag{124}$$

$$(1 - c(1 - (\delta^0)^2))^{\frac{T}{2}} \leq \frac{\epsilon}{\delta^0} \tag{125}$$

$$\frac{T}{2} \log\left(1 - c(1 - (\delta^0)^2)\right) \leq \log\left(\frac{\epsilon}{\delta^0}\right) \tag{126}$$

$$\tag{127}$$

We proceed such that

$$T \geq \frac{2 \log(\frac{\epsilon}{\delta^0})}{\log\left(1 - c(1 - (\delta^0)^2)\right)} \tag{128}$$

$$> \frac{2 \log(\frac{\epsilon}{\delta^0})}{-c(1 - (\delta^0)^2)} \tag{129}$$

$$= \frac{2}{c(1 - (\delta^0)^2)} \log\left(\frac{\delta^0}{\epsilon}\right) \tag{130}$$

where (129) follows by using $\log(1-x) < -x$ for $|x| < 1$. Thus, with probability at least $1 - 2Tq^{-10}$, $\delta^T = \sin\theta(\mathbf{a}^T, \mathbf{a}^*) \leq \epsilon$.

Convergence to the target model. We now aim to upper bound $\|\mathbf{a}^T (\mathbf{b}^{T+1})^\top - \mathbf{a}^* (\mathbf{b}^*)^\top\|$. Recall that $(\mathbf{g}^T)^\top = (\mathbf{a}^T)^\top \mathbf{a}^* {\mathbf{b}^*}^\top$ and $\delta^T = \|(\mathbf{I}_d - \mathbf{a}^T (\mathbf{a}^T)^\top)\mathbf{a}^*\|$, we have

$$\|\mathbf{a}^T (\mathbf{b}^{T+1})^\top - \mathbf{a}^* {\mathbf{b}^*}^\top\| = \|\mathbf{a}^T (\mathbf{b}^{T+1})^\top - \mathbf{a}^T (\mathbf{g}^T)^\top + \mathbf{a}^T (\mathbf{g}^T)^\top - \mathbf{a}^* {\mathbf{b}^*}^\top\| \tag{131}$$

$$\leq \|\mathbf{a}^T (\mathbf{b}^{T+1})^\top - \mathbf{a}^T (\mathbf{g}^T)^\top\| + \|\mathbf{a}^T (\mathbf{g}^T)^\top - \mathbf{a}^* {\mathbf{b}^*}^\top\| \tag{132}$$

$$= \|\mathbf{a}^T (\mathbf{b}^{T+1} - \mathbf{g}^T)^\top\| + \|(\mathbf{a}^T (\mathbf{a}^T)^\top - \mathbf{I}_d)\mathbf{a}^* {\mathbf{b}^*}^\top\| \tag{133}$$

$$= \|\mathbf{a}^T\| \|\mathbf{b}^{T+1} - \mathbf{g}^T\| + \|(\mathbf{I}_d - \mathbf{a}^T (\mathbf{a}^T)^\top)\mathbf{a}^*\| \|\mathbf{b}^*\| \tag{134}$$

$$\leq \epsilon' \delta^T \|\mathbf{b}^*\| + \delta^T \|\mathbf{b}^*\| \tag{135}$$

$$= (1 + \epsilon')\epsilon \|\mathbf{b}^*\| \tag{136}$$

$$= (1 + \epsilon')\epsilon \|\mathbf{a}^* {\mathbf{b}^*}^\top\| \tag{137}$$

where (135) is by Lemma A3.9 and the fact that $\|\mathbf{a}^T\| = 1$, and (137) is due to the fact that $\|\mathbf{x}\mathbf{y}^\top\| = \|\mathbf{x}\|\|\mathbf{y}\|$ and $\|\mathbf{a}^*\| = 1$. □

A3.4.1 Proof of Proposition 5.5

Proof. We start by fixing $\mathbf{a}^0$ and updating $\mathbf{b}_i$ to minimize the objective. Let $\mathbf{a} = \mathbf{a}^0$. We obtain

$$\mathbf{b}_i^\top = \frac{\mathbf{a}^\top \mathbf{X}_i^\top \mathbf{X}_i \mathbf{a}^* \mathbf{b}^{*\top}}{\mathbf{a}^\top \mathbf{X}_i^\top \mathbf{X}_i \mathbf{a}} \tag{138}$$

$$(\mathbf{b}^{FFA})^\top = \frac{1}{N} \sum_{i=1}^N \frac{\mathbf{a}^\top \mathbf{X}_i^\top \mathbf{X}_i \mathbf{a}^* \mathbf{b}^{*\top}}{\mathbf{a}^\top \mathbf{X}_i^\top \mathbf{X}_i \mathbf{a}} \tag{139}$$

let $\bar{\mathbf{b}} = \mathbf{b}^{FFA}$. We aim to compute the expected value of $\frac{1}{N}\sum_{i=1}^N \frac{1}{m}\|\mathbf{X}_i \mathbf{a}^* \mathbf{b}^{*\top} - \mathbf{X}_i \mathbf{a} \bar{\mathbf{b}}^\top\|^2$ where the expectation is over all the randomness in the $\mathbf{X}_i$. We define

$$s_i = \frac{\mathbf{a}^\top \mathbf{X}_i^\top \mathbf{X}_i \mathbf{a}^*}{\mathbf{a}^\top \mathbf{X}_i^\top \mathbf{X}_i \mathbf{a}} = \frac{(\mathbf{X_i a})^\top (\mathbf{X_i a}^*)}{\|\mathbf{X}_i \mathbf{a}\|^2} \tag{140}$$

so that $\bar{\mathbf{b}} = \frac{1}{N}\sum_{i=1}^N s_i \mathbf{b}^* = \bar{s}\mathbf{b}^*$. For each i, the norm becomes

$$\begin{aligned} \|\mathbf{X}_i \mathbf{a}^* \mathbf{b}^{*\top} - \mathbf{X}_i \mathbf{a} \bar{\mathbf{b}}^\top\|^2 &= \|(\mathbf{X}_i \mathbf{a}^* - \bar{s}\mathbf{X}_i \mathbf{a})\mathbf{b}^{*\top}\|^2 & (141) \\ &= \|\mathbf{X}_i \mathbf{a}^* - \bar{s}\mathbf{X}_i \mathbf{a}\|^2 \|\mathbf{b}^*\|^2 & (142) \end{aligned}$$

using the fact that $\|\mathbf{u}\mathbf{v}^\top\|^2 = \|\mathbf{u}\|^2\|\mathbf{v}\|^2$ for vectors $\mathbf{u}$ and $\mathbf{v}$. Therefore, $\mathbb{E}[\frac{1}{N}\sum_{i=1}^N \frac{1}{m}\|\mathbf{X}_i \mathbf{a}^* \mathbf{b}^{*\top} - \mathbf{X}_i \mathbf{a} \bar{\mathbf{b}}^\top\|^2]$ is reduced to $\mathbb{E}[\frac{1}{N}\sum_{i=1}^N \frac{1}{m}\|\mathbf{X}_i \mathbf{a}^* - \mathbf{X}_i \mathbf{a}\|^2] \cdot \|\mathbf{b}^*\|^2$.

Since each entry of $\mathbf{X}_i$ is independently and identically distributed according to a standard Gaussian distribution, both $\mathbf{a}^*$ and $\mathbf{a}$ are unit vectors, the vectors $\mathbf{X}_i \mathbf{a}^*$ and $\mathbf{X}_i \mathbf{a}$ are $\mathcal{N}(0, \mathbf{I}_m)$. The cross-covariance is $\alpha \mathbf{I}_m$ where $\alpha = \mathbf{a}^\top \mathbf{a}^*$.

By linearity, we can show that $\frac{1}{N}\sum_{i=1}^N \frac{1}{m}\|\mathbf{X}_i \mathbf{a}^* - \bar{s}\mathbf{X}_i \mathbf{a}\|^2$ has the same expectation as $\frac{1}{m}\|\mathbf{X}_1 \mathbf{a}^* - \bar{s}\mathbf{X}_1 \mathbf{a}\|^2$ because all $(\mathbf{X}_i \mathbf{a}^*, \mathbf{X}_i \mathbf{a})$ are i.i.d. pairs. Let $z_1 = \frac{s_1}{N}$ and $z_2 = \frac{s_2 + \cdots + s_N}{N}$, we have $\|\mathbf{X}_1 \mathbf{a}^* - z_1 \mathbf{X}_1 \mathbf{a} - z_2 \mathbf{X}_1 \mathbf{a}\|^2$. Let $\mathbf{v} = \mathbf{X}_1 \mathbf{a}^*, \mathbf{u}_1 = z_1 \mathbf{X}_1 \mathbf{a}$ and $\mathbf{u}_2 = z_2 \mathbf{X}_1 \mathbf{a}$. Thus,

$$\begin{aligned} \|\mathbf{X}_1 \mathbf{a}^* - z_1 \mathbf{X}_1 \mathbf{a} - z_2 \mathbf{X}_1 \mathbf{a}\|^2 &= \|\mathbf{v} - \mathbf{u}_1 - \mathbf{u}_2\|^2 & (143) \\ &= \mathbf{v}^\top \mathbf{v} + \mathbf{u}_1^\top \mathbf{u}_1 + \mathbf{u}_2^\top \mathbf{u}_2 - 2\mathbf{v}^\top \mathbf{u}_1 - 2\mathbf{v}^\top \mathbf{u}_2 + 2\mathbf{u}_1^\top \mathbf{u}_2 & (144) \end{aligned}$$

Now we compute the expectation term by term.

Expected value of $\mathbf{v}^\top \mathbf{v}$ We have $\mathbb{E}[\mathbf{v}^\top \mathbf{v}] = \mathbb{E}[\|\mathbf{X}_1 \mathbf{a}^*\|^2] = m$.

Expected value of $\mathbf{u}_1^\top \mathbf{u}_1$ We have

$$\begin{aligned} \mathbf{u}_1^\top \mathbf{u}_1 &= z_1^2 \|\mathbf{X}_1 \mathbf{a}\|^2 & (145) \\ &= \frac{s_1^2}{N^2}\|\mathbf{X}_1 \mathbf{a}\|^2 & (146) \\ &= \frac{1}{N^2}\frac{((\mathbf{X}_1 \mathbf{a})^\top (\mathbf{X}_1 \mathbf{a}^*))^2}{\|\mathbf{X}_1 \mathbf{a}\|^4}\|\mathbf{X}_1 \mathbf{a}\|^2 & (147) \\ &= \frac{1}{N^2}\frac{((\mathbf{X}_1 \mathbf{a})^\top (\mathbf{X}_1 \mathbf{a}^*))^2}{\|\mathbf{X}_1 \mathbf{a}\|^2} & (148) \end{aligned}$$

Note that $(\mathbf{X}_1 \mathbf{a}^*, \mathbf{X}_1 \mathbf{a})$ is a correlated Gaussian pair with correlation $\alpha = \mathbf{a}^\top \mathbf{a}^*$. Without loss of generality, we assume $\mathbf{a} = \mathbf{e}_1$ thus $\mathbf{a}^* = \alpha \mathbf{e}_1 + \sqrt{1-\alpha^2}\mathbf{e}_2$, where $\mathbf{e}_1$ and $\mathbf{e}_2$ are standard basis vectors in $\mathbb{R}^d$. So we can get $\mathbf{X}_1 \mathbf{a} = \mathbf{X}_1 \mathbf{e}_1 = \mathbf{x}_{1,1}$, where $\mathbf{x}_{1,1}$ denotes the first column of $\mathbf{X}_1$. Accordingly $\mathbf{X}_1 \mathbf{a}^* = \alpha \mathbf{X}_1 \mathbf{e}_1 + \sqrt{1-\alpha^2}\mathbf{X}_1 \mathbf{e}_2 = \alpha \mathbf{x}_{1,1} + \sqrt{1-\alpha^2}\mathbf{x}_{1,2}$ where $\mathbf{x}_{1,2}$ denotes the second column of $\mathbf{X}_1$. Therefore (148) can be written as $\frac{1}{N^2}\frac{(\mathbf{x}_{1,1}^\top(\alpha \mathbf{x}_{1,1} + \beta \mathbf{x}_{1,2}))^2}{\|\mathbf{x}_{1,1}\|^2}$. Now we take expectation of it.

$$\mathbb{E}\left[\frac{1}{N^2}\frac{((\mathbf{X}_1 \mathbf{a})^\top (\mathbf{X}_1 \mathbf{a}^*))^2}{\|\mathbf{X}_1 \mathbf{a}\|^2}\right] = \mathbb{E}\left[\frac{1}{N^2}\frac{(\mathbf{x}_{1,1}^\top(\alpha \mathbf{x}_{1,1} + \beta \mathbf{x}_{1,2}))^2}{\|\mathbf{x}_{1,1}\|^2}\right] = \frac{1}{N^2}\mathbb{E}\left[\frac{(\mathbf{x}_{1,1}^\top(\alpha \mathbf{x}_{1,1} + \beta \mathbf{x}_{1,2}))^2}{\|\mathbf{x}_{1,1}\|^2}\right] \tag{149}$$

Let $r_1 = \|\mathbf{x}_{1,1}\|^2$ and $r_2 = \mathbf{x}_{1,1}^\top \mathbf{x}_{1,2}$. We have

$$\mathbb{E}\left[\frac{(\mathbf{x}_{1,1}^\top(\alpha \mathbf{x}_{1,1} + \beta \mathbf{x}_{1,2}))^2}{\|\mathbf{x}_{1,1}\|^2}\right] = \mathbb{E}\left[\frac{(\alpha r_1 + \beta r_2)^2}{r_1}\right] \tag{150}$$

$$= \mathbb{E}\left[\frac{\alpha^2 r_1^2 + \beta^2 r_2^2 + 2\alpha\beta r_1 r_2}{r_1}\right] \tag{151}$$

$$= \mathbb{E}\left[\alpha^2 r_1\right] + \mathbb{E}\left[\frac{\beta^2 r_2^2}{r_1}\right] + \mathbb{E}\left[2\alpha\beta r_2\right] \tag{152}$$

where $\mathbb{E}\left[\alpha^2 r_1\right] = \alpha^2 \mathbb{E}\left[\|\mathbf{x}_{1,1}\|^2\right] = \alpha^2 m$, and $\mathbb{E}\left[2\alpha\beta r_2\right] = 2\alpha\beta \mathbb{E}\left[r_2\right] = 2\alpha\beta \mathbb{E}\left[\mathbf{x}_{1,1}^\top \mathbf{x}_{1,2}\right] = 0$ because $\mathbf{x}_{1,1}$ and $\mathbf{x}_{1,2}$ are independent standard Gaussian vectors. Then we analyze $\mathbb{E}\left[\frac{\beta^2 r_2^2}{r_1}\right] = \beta^2 \mathbb{E}\left[\frac{r_2^2}{r_1}\right]$. Condition on $\mathbf{x}_{1,1}$,

$$\mathbb{E}\left[r_2 | \mathbf{x}_{1,1}\right] = \mathbb{E}\left[\mathbf{x}_{1,1}^\top \mathbf{x}_{1,2} | \mathbf{x}_{1,1}\right] = \mathbf{x}_{1,1}^\top \mathbb{E}\left[\mathbf{x}_{1,2}\right] = 0 \tag{153}$$

and $\mathrm{Var}(r_2 | \mathbf{x}_{1,1}) = \|\mathbf{x}_{1,1}\|^2 = r_1$, thus

$$r_2 | \mathbf{x}_{1,1} = \mathbf{x}_{1,1}^\top \mathbf{x}_{1,2} | \mathbf{x}_{1,1} \sim \mathcal{N}(0, r_1) \tag{154}$$

Then we obtain

$$\mathbb{E}\left[r_2^2 | \mathbf{x}_{1,1}\right] = r_1 \tag{155}$$

Therefore $\mathbb{E}\left[\frac{r_2^2}{r_1} | \mathbf{x}_{1,1}\right] = \frac{\mathbb{E}\left[r_2^2 | \mathbf{x}_{1,1}\right]}{r_1} = 1$. We take total expectation $\mathbb{E}\left[\frac{r_2^2}{r_1}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{r_2^2}{r_1} | \mathbf{x}_{1,1}\right]\right] = 1$. Summarizing,

$$\mathbb{E}\left[\frac{((\mathbf{X}_1 \mathbf{a})^\top (\mathbf{X}_1 \mathbf{a}^*))^2}{\|\mathbf{X}_1 \mathbf{a}\|^2}\right] = \mathbb{E}\left[\frac{(\alpha r_1 + \beta r_2)^2}{r_1}\right] \tag{156}$$

$$= \mathbb{E}\left[\alpha^2 r_1\right] + \mathbb{E}\left[\frac{\beta^2 r_2^2}{r_1}\right] + \mathbb{E}\left[2\alpha\beta r_2\right] \tag{157}$$

$$= \alpha^2 m + \beta^2 \tag{158}$$

$$\mathbb{E}\left[\mathbf{u}_1^\top \mathbf{u}_1\right] = \frac{1}{N^2} \mathbb{E}\left[\frac{((\mathbf{X}_1 \mathbf{a})^\top (\mathbf{X}_1 \mathbf{a}^*))^2}{\|\mathbf{X}_1 \mathbf{a}\|^2}\right] \tag{159}$$

$$= \frac{\alpha^2 m + (1 - \alpha^2)}{N^2} \tag{160}$$

Expected value of $\mathbf{u}_2^\top \mathbf{u}_2$ We have $\mathbf{u}_2^\top \mathbf{u}_2 = z_2^2 \|\mathbf{X}_1 \mathbf{a}\|^2$ where $z_2 = \frac{s_2 + \cdots + s_N}{N}$ is independent of pair $(\mathbf{X}_1 \mathbf{a}^*, \mathbf{X}_1 \mathbf{a})$. To compute $\mathbb{E}\left[z_2^2 \|\mathbf{X}_1 \mathbf{a}\|^2\right]$, first we condition on z_2 to obtain,

$$\mathbb{E}\left[z_2^2 \|\mathbf{X}_1 \mathbf{a}\|^2 | z_2\right] = z_2^2 \mathbb{E}\left[\|\mathbf{X}_1 \mathbf{a}\|^2\right] = z_2^2 m \tag{161}$$

Then we take total expectation $\mathbb{E}\left[z_2^2 \|\mathbf{X}_1 \mathbf{a}\|^2\right] = \mathbb{E}\left[\mathbb{E}\left[z_2^2 \|\mathbf{X}_1 \mathbf{a}\|^2 | z_2\right]\right] = \mathbb{E}\left[z_2^2 m\right] = m \mathbb{E}\left[z_2^2\right]$.

$$\mathbb{E}\left[z_2^2\right] = \mathbb{E}\left[\frac{(s_2 + \cdots + s_N)^2}{N^2}\right] \tag{162}$$

$$= \frac{1}{N^2} \mathbb{E}\left[\sum_{i=2}^N s_i^2 + \sum_{\substack{i=1, j=1 \\ i \neq j}}^N s_i s_j\right] \tag{163}$$

$$= \frac{1}{N^2} \left(\sum_{i=2}^N \mathbb{E}\left[s_i^2\right] + \sum_{\substack{i=1, j=1 \\ i \neq j}}^N \mathbb{E}\left[s_i s_j\right]\right) \tag{164}$$

Write $s_i = \frac{(\mathbf{X_i a})^\top (\mathbf{X_i a}^*)}{\|\mathbf{X}_i \mathbf{a}\|^2}$. Without loss of generality, we assume $\mathbf{a} = \mathbf{e}_1$ thus $\mathbf{a}^* = \alpha \mathbf{e}_1 + \sqrt{1-\alpha^2}\mathbf{e}_2$, where $\mathbf{e}_1$ and $\mathbf{e}_2$ are standard basis vectors in $\mathbb{R}^d$. Thus, we have $\mathbf{X}_i \mathbf{a} = \mathbf{X}_i \mathbf{e}_1 = \mathbf{x}_{i,1}$, where $\mathbf{x}_{i,1}$ represents the first column of $\mathbf{X}_i$. Similarly, $\mathbf{X}_i \mathbf{a}^* = \alpha \mathbf{X}_i \mathbf{e}_1 + \sqrt{1-\alpha^2}\mathbf{X}_i \mathbf{e}_2 = \alpha \mathbf{x}_{i,1} + \sqrt{1-\alpha^2}\mathbf{x}_{i,2}$, where $\mathbf{x}_{i,2}$ denotes the second column of $\mathbf{X}_i$.

Hence,

$$(\mathbf{X_i a})^\top (\mathbf{X_i a}^*) = \mathbf{x}_{i,1}^\top (\alpha \mathbf{x}_{i,1} + \sqrt{1-\alpha^2}\mathbf{x}_{i,2}) = \alpha \|\mathbf{x}_{i,1}\|^2 + \sqrt{1-\alpha^2}(\mathbf{x}_{i,1}^\top \mathbf{x}_{i,2}) \tag{165}$$

With $\|\mathbf{X_i a}\|^2 = \|\mathbf{x}_{i,1}\|^2$, we have

$$s_i = \frac{\alpha \|\mathbf{x}_{i,1}\|^2 + \sqrt{1-\alpha^2}(\mathbf{x}_{i,1}^\top \mathbf{x}_{i,2})}{\|\mathbf{x}_{i,1}\|^2} = \alpha + \sqrt{1-\alpha^2}\frac{\mathbf{x}_{i,1}^\top \mathbf{x}_{i,2}}{\|\mathbf{x}_{i,1}\|^2} \tag{166}$$

Let $R = \frac{\mathbf{x}_{i,1}^\top \mathbf{x}_{i,2}}{\|\mathbf{x}_{i,1}\|^2}$. Then

$$\begin{aligned}
\mathbb{E}\left[s_i^2\right] &= \mathbb{E}\left[\left(\alpha + \sqrt{1-\alpha^2}R\right)^2\right] && (167)\\
&= \mathbb{E}\left[\alpha^2 + (1-\alpha^2)R^2 + 2\alpha\sqrt{1-\alpha^2}R\right] && (168)\\
&= \alpha^2 + (1-\alpha^2)\mathbb{E}\left[R^2\right] + 2\alpha\sqrt{1-\alpha^2}\mathbb{E}\left[R\right] && (169)
\end{aligned}$$

For $\mathbb{E}\left[R\right] = \mathbb{E}\left[\frac{\mathbf{x}_{i,1}^\top \mathbf{x}_{i,2}}{\|\mathbf{x}_{i,1}\|^2}\right]$, similarly as in (154), $\mathbf{x}_{i,1}^\top \mathbf{x}_{i,2} | \mathbf{x}_{i,1} \sim \mathcal{N}(0, \|\mathbf{x}_{i,1}\|^2)$, thus $\frac{\mathbf{x}_{i,1}^\top \mathbf{x}_{i,2}}{\|\mathbf{x}_{i,1}\|^2} | \mathbf{x}_{i,1} \sim \mathcal{N}(0, \frac{1}{\|\mathbf{x}_{i,1}\|^2})$, then

$$\mathbb{E}\left[R\right] = \mathbb{E}\left[\mathbb{E}\left[R | \mathbf{x}_{i,1}\right]\right] = 0 \tag{170}$$

For $\mathbb{E}\left[R^2\right]$, since $\mathbf{x}_{i,1}^\top \mathbf{x}_{i,2} | \mathbf{x}_{i,1} \sim \mathcal{N}(0, \|\mathbf{x}_{i,1}\|^2)$, so $\mathbb{E}\left[(\mathbf{x}_{i,1}^\top \mathbf{x}_{i,2})^2 | \mathbf{x}_{i,1}\right] = \|\mathbf{x}_{i,1}\|^2$. Thus, with $R^2 = \frac{(\mathbf{x}_{i,1}^\top \mathbf{x}_{i,2})^2}{\|\mathbf{x}_{i,1}\|^4}$,

$$\begin{aligned}
\mathbb{E}\left[R^2 | \mathbf{x}_{i,1}\right] &= \frac{\mathbb{E}\left[(\mathbf{x}_{i,1}^\top \mathbf{x}_{i,2})^2 | \mathbf{x}_{i,1}\right]}{\|\mathbf{x}_{i,1}\|^4} = \frac{1}{\|\mathbf{x}_{i,1}\|^2} && (171)\\
\mathbb{E}\left[R^2\right] &= \mathbb{E}\left[\frac{1}{\|\mathbf{x}_{i,1}\|^2}\right] && (172)
\end{aligned}$$

For a m-dimensional standard Gaussian vector, $\|\mathbf{x}_{i,1}\|^2$ follows a chi-squared distribution with m degrees of freedom. Therefore, $\mathbb{E}\left[R^2\right] = \frac{1}{m-2}$. (169) becomes

$$\begin{aligned}
\mathbb{E}\left[s_i^2\right] &= \alpha^2 + (1-\alpha^2)\mathbb{E}\left[R^2\right] + 2\alpha\sqrt{1-\alpha^2}\mathbb{E}\left[R\right] && (173)\\
&= \alpha^2 + (1-\alpha^2)\frac{1}{m-2} && (174)
\end{aligned}$$

Now we compute $\mathbb{E}\left[s_i s_j\right]$ for $i \neq j$. By independence of s_i and s_j, $\mathbb{E}\left[s_i s_j\right] = \mathbb{E}\left[s_i\right] \cdot \mathbb{E}\left[s_j\right] = \mathbb{E}\left[s_i\right]^2$. Take expectation of (166),

$$\begin{aligned}
\mathbb{E}\left[s_i\right] &= \mathbb{E}\left[\alpha + \sqrt{1-\alpha^2}\frac{\mathbf{x}_{i,1}^\top \mathbf{x}_{i,2}}{\|\mathbf{x}_{i,1}\|^2}\right] && (175)\\
&= \alpha + \sqrt{1-\alpha^2}\mathbb{E}\left[\frac{\mathbf{x}_{i,1}^\top \mathbf{x}_{i,2}}{\|\mathbf{x}_{i,1}\|^2}\right] && (176)\\
&= \alpha + \sqrt{1-\alpha^2}\mathbb{E}\left[R\right] && (177)\\
&= \alpha && (178)
\end{aligned}$$

Hence, $\mathbb{E}[s_i s_j] = \alpha^2$. Summarizing,

$$\mathbb{E}\left[\mathbf{u}_2^\top \mathbf{u}_2\right] = m\mathbb{E}\left[z_2^2\right] \tag{179}$$

$$= \frac{m}{N^2}\left((N-1)\mathbb{E}\left[s_i^2\right] + (N-1)(N-2)\mathbb{E}[s_i s_j]\right) \tag{180}$$

$$= \frac{m}{N^2}\left((N-1)(\alpha^2 + (1-\alpha^2)\frac{1}{m-2}) + (N-1)(N-2)\alpha^2\right) \tag{181}$$

$$= \frac{m}{N^2}\left[(N-1)^2\alpha^2 + (N-1)\frac{1-\alpha^2}{m-2}\right] \tag{182}$$

Expected value of $\mathbf{v}^\top \mathbf{u}_1$ We have $\mathbf{v}^\top \mathbf{u}_1 = z_1(\mathbf{X}_1\mathbf{a}^*)^\top(\mathbf{X}_1\mathbf{a}) = \frac{s_1}{N}(\mathbf{X}_1\mathbf{a}^*)^\top(\mathbf{X}_1\mathbf{a})$. We factor out $\frac{1}{N}$, $\mathbb{E}\left[\mathbf{v}^\top \mathbf{u}_1\right] = \frac{1}{N}\mathbb{E}\left[s_1(\mathbf{X}_1\mathbf{a}^*)^\top(\mathbf{X}_1\mathbf{a})\right]$. By (158),

$$\mathbb{E}\left[s_1(\mathbf{X}_1\mathbf{a}^*)^\top(\mathbf{X}_1\mathbf{a})\right] = \mathbb{E}\left[\frac{((\mathbf{X}_1\mathbf{a})^\top(\mathbf{X}_1\mathbf{a}^*))^2}{\|\mathbf{X}_1\mathbf{a}\|^2}\right] \tag{183}$$

$$= \alpha^2 m + (1-\alpha^2) \tag{184}$$

Then

$$\mathbb{E}\left[\mathbf{v}^\top \mathbf{u}_1\right] = \frac{\alpha^2 m + (1-\alpha^2)}{N} \tag{185}$$

Expected value of $\mathbf{v}^\top \mathbf{u}_2$ We have $\mathbf{v}^\top \mathbf{u}_2 = z_2(\mathbf{X}_1\mathbf{a}^*)^\top(\mathbf{X}_1\mathbf{a})$. Condition on z_2 which is independent of $(\mathbf{X}_1\mathbf{a}^*)^\top(\mathbf{X}_1\mathbf{a})$, we obtain

$$\mathbb{E}\left[z_2(\mathbf{X}_1\mathbf{a}^*)^\top(\mathbf{X}_1\mathbf{a})|z_2\right] = z_2\mathbb{E}\left[(\mathbf{X}_1\mathbf{a}^*)^\top(\mathbf{X}_1\mathbf{a})\right] \tag{186}$$

Still we assume $\mathbf{a} = \mathbf{e}_1$ thus $\mathbf{a}^* = \alpha\mathbf{e}_1 + \sqrt{1-\alpha^2}\mathbf{e}_2$, where $\mathbf{e}_1$ and $\mathbf{e}_2$ are standard basis vectors in $\mathbb{R}^d$. With $\mathbf{X}_1\mathbf{a} = \mathbf{X}_1\mathbf{e}_1 = \mathbf{x}_{1,1}$, where $\mathbf{x}_{1,1}$ denotes the first column of $\mathbf{X}_1$, and $\mathbf{X}_1\mathbf{a}^* = \alpha\mathbf{X}_1\mathbf{e}_1 + \sqrt{1-\alpha^2}\mathbf{X}_1\mathbf{e}_2 = \alpha\mathbf{x}_{1,1} + \sqrt{1-\alpha^2}\mathbf{x}_{1,2}$ where $\mathbf{x}_{1,2}$ denotes the second column of $\mathbf{X}_1$, using (165),

$$\mathbb{E}\left[(\mathbf{X}_1\mathbf{a}^*)^\top(\mathbf{X}_1\mathbf{a})\right] = \mathbb{E}\left[\alpha\|\mathbf{x}_{1,1}\|^2 + \sqrt{1-\alpha^2}(\mathbf{x}_{1,1}^\top\mathbf{x}_{1,2})\right] \tag{187}$$

$$= \alpha\mathbb{E}\left[\|\mathbf{x}_{1,1}\|^2\right] + z_2\sqrt{1-\alpha^2}\mathbb{E}\left[\mathbf{x}_{1,1}^\top\mathbf{x}_{1,2}\right] \tag{188}$$

$$= \alpha m \tag{189}$$

Thus $z_2\mathbb{E}\left[(\mathbf{X}_1\mathbf{a}^*)^\top(\mathbf{X}_1\mathbf{a})\right] = z_2\alpha m$, Then we take total expectation

$$\mathbb{E}\left[\mathbb{E}\left[z_2(\mathbf{X}_1\mathbf{a}^*)^\top(\mathbf{X}_1\mathbf{a})|z_2\right]\right] = \mathbb{E}[z_2\alpha m] \tag{190}$$

$$= \alpha m\mathbb{E}[z_2] \tag{191}$$

where $z_2 = \frac{s_2+\cdots+s_N}{N}$. Therefore,

$$\alpha m\mathbb{E}[z_2] = \frac{\alpha m}{N}\sum_{i=2}^{N}\mathbb{E}[s_i] = \frac{m}{N}(N-1)\alpha^2 \tag{192}$$

where (192) follows by $\mathbb{E}[s_i] = \alpha$. Summarizing, we obtain $\mathbb{E}\left[\mathbf{v}^\top\mathbf{u}_2\right] = \frac{m}{N}(N-1)\alpha^2$.

Expected value of $\mathbf{u}_1^\top \mathbf{u}_2$ We have $\mathbf{u}_1^\top\mathbf{u}_2 = z_1 z_2\|\mathbf{X}_1\mathbf{a}\|^2$. By definition of z_1 and z_2, we obtain

$$z_1 z_2\|\mathbf{X}_1\mathbf{a}\|^2 = \frac{1}{N^2}((\mathbf{X}_1\mathbf{a}^*)^\top(\mathbf{X}_1\mathbf{a}))\sum_{i=2}^{N} s_i \tag{193}$$

Since $(\mathbf{X}_1\mathbf{a}^*)^\top(\mathbf{X}_1\mathbf{a})$ depends only on $\mathbf{X}_1$, $\sum_{i=2}^{N} s_i$ is independent of $\mathbf{X}_1$, we obtain

$$\mathbb{E}\left[\frac{1}{N^2}((\mathbf{X}_1\mathbf{a}^*)^\top(\mathbf{X}_1\mathbf{a}))\sum_{i=2}^{N} s_i\right] = \frac{1}{N^2}\mathbb{E}\left[(\mathbf{X}_1\mathbf{a}^*)^\top(\mathbf{X}_1\mathbf{a})\right]\cdot\mathbb{E}\left[\sum_{i=2}^{N} s_i\right] \tag{194}$$

$$= \frac{1}{N^2}\mathbb{E}\left[(\mathbf{X}_1\mathbf{a}^*)^\top(\mathbf{X}_1\mathbf{a})\right]\cdot(N-1)\mathbb{E}[s_i] \tag{195}$$

$$= \frac{(N-1)m\alpha^2}{N^2} \tag{196}$$

where (196) follows by $\mathbb{E}\left[(\mathbf{X}_1\mathbf{a}^*)^\top(\mathbf{X}_1\mathbf{a})\right] = \alpha m$ and $\mathbb{E}\left[s_i\right] = \alpha$.

Combining (160), (182),(185),(192),(196) and (144),

$$\frac{1}{m}\|\mathbf{X}_1\mathbf{a}^* - \bar{s}\mathbf{X}_1\mathbf{a}\|^2 = \frac{1}{m}(\mathbf{v}^\top\mathbf{v} + \mathbf{u}_1^\top\mathbf{u}_1 + \mathbf{u}_2^\top\mathbf{u}_2 - 2\mathbf{v}^\top\mathbf{u}_1 - 2\mathbf{v}^\top\mathbf{u}_2 + 2\mathbf{u}_1^\top\mathbf{u}_2) \tag{197}$$

$$= (1-\alpha^2)\left[1 + \frac{N(4-m)-2}{N^2m(m-2)}\right] \tag{198}$$

$$= (\delta^0)^2(1+\tilde{c}) \tag{199}$$

where δ^0 is the angle distance between $\mathbf{a}$ and $\mathbf{a}^*$. The quantity $\tilde{c} = \frac{N(4-m)-2}{N^2m(m-2)} = O(\frac{1}{Nm})$ as N and m approach infinity. Therefore,

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\frac{1}{m}\|\mathbf{X}_i\mathbf{a}^*\mathbf{b}^{*\top} - \mathbf{X}_i\mathbf{a}\bar{\mathbf{b}}^\top\|^2\right] = (1+\tilde{c})(\delta^0)^2\|\mathbf{b}^*\|^2 \tag{200}$$

□

A3.5 Heterogeneous Case

Consider a federated setting with N clients, each with the following local linear model

$$f_i(\mathbf{X}_i) = \mathbf{X}_i\mathbf{a}\mathbf{b}^\top \tag{201}$$

where $\mathbf{a} \in \mathbb{R}^d$ is a unit vector and $\mathbf{b} \in \mathbb{R}^d$ are the LoRA weights corresponding to rank $r = 1$. In this setting, we model the local data of i-th client such that $\mathbf{Y}_i = \mathbf{X}_i\mathbf{a}^*\mathbf{b}_i^{*\top}$ for some ground truth LoRA weights $\mathbf{a}^* \in \mathbb{R}^d$, which is a unit vector, and local $\mathbf{b}_i^* \in \mathbb{R}^d$. We consider the following objective

$$\min_{\mathbf{a}\in\mathbb{R}^d,\mathbf{b}\in\mathbb{R}^d} \frac{1}{N}\sum_{i=1}^{N} l_i(\mathbf{a},\mathbf{b}) \tag{202}$$

We consider the local population loss $l_i(\mathbf{a},\mathbf{b}) = \|\mathbf{a}^*\mathbf{b}_i^{*\top} - \mathbf{a}\mathbf{b}^\top\|^2$.

We aim to learn a shared model $(\mathbf{a},\mathbf{b})$ for all the clients. It is straightforward to observe that $(\mathbf{a}',\mathbf{b}')$ is a global minimizer of if and only if $\mathbf{a}'\mathbf{b}'^\top = \mathbf{a}^*\bar{\mathbf{b}}^*$, where $\bar{\mathbf{b}}^* = \frac{1}{N}\sum_{i=1}^{N}\mathbf{b}_i^*$. The solution is unique and satisfies $\mathbf{a}' = \mathbf{a}^*$ and $\mathbf{b}' = \bar{\mathbf{b}}^*$. With this global minimizer, we obtain the corresponding minimum global error of $\frac{1}{N}\sum_{i=1}^{N}\|\mathbf{a}^*(\mathbf{b}_i^* - \bar{\mathbf{b}}^*)^\top\|^2$.

We aim to show that the training procedure described in Algorithm 2 learns the global minimizer $(\mathbf{a}^*, \bar{\mathbf{b}}^*)$. First, we make typical assumption and definition.

Assumption A3.12. There exists $L_{max} < \infty$ (known a priori), s.t. $\|\bar{\mathbf{b}}^*\| \le L_{max}$.

Definition A3.13. (Client variance) For $\gamma > 0$, we define $\gamma^2 := \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{b}_i^* - \bar{\mathbf{b}}^*\|^2$, where $\bar{\mathbf{b}}^* = \frac{1}{N}\sum_{i=1}^{N}\mathbf{b}_i^*$.

Theorem A3.14. *(Convergence of RoLoRA for linear regressor in heterogeneous setting) Let* $\delta^t = \|(\mathbf{I}_d - \mathbf{a}^*\mathbf{a}^{*\top})\mathbf{a}^t\|$ *be the angle distance between* $\mathbf{a}^*$ *and* $\mathbf{a}^t$ *of* t*-th iteration. Suppose we are in the setting described in Section A3.5 and apply Algorithm 2 for optimization. Given a random initial* $\mathbf{a}^0$*, an initial angle distance* $\delta_0 \in (0,1)$*, we set the step size* $\eta \le \frac{1}{2L_{\max}^2}$ *and the number of iterations* $T \ge \frac{1}{c(1-(\delta^0)^2)}\log(\frac{\delta^0}{\epsilon})$ *for* $c \in (0,1)$*. Under these conditions, we achieve the following*

$$\sin\theta(\mathbf{a}^T,\mathbf{a}^*) \le \epsilon, \text{ and } \|\mathbf{a}^T(\mathbf{b}^{T+1})^\top - \mathbf{a}^*(\bar{\mathbf{b}}^*)^\top\| \le \epsilon\|\mathbf{a}^*(\bar{\mathbf{b}}^*)^\top\|$$

which we refer to as ϵ*-accurate recovery of the global minimizer.*

Theorem A3.14 follows by recursively applying Lemma A3.16 for T iterations. We start by computing the update rule for $\mathbf{a}$ as in Lemma A3.15. Using Lemma A3.15, we analyze the convergence of $\mathbf{a}$ in Lemma A3.16. We also show the global error that can be achieved by FFA-LoRA within this setting in Proposition A3.17.

Lemma A3.15. *(Update for* $\mathbf{a}$*) In RoLoRA for linear regressor, the update for* $\mathbf{a}$ *and* b *in each iteration is:*

$$\mathbf{b}^{t+1} = \bar{\mathbf{b}} = \bar{\mathbf{b}}^* {\mathbf{a}^*}^\top \mathbf{a}^t \tag{203}$$

$$\mathbf{a}^{t+1} = \hat{\mathbf{a}} = \frac{\mathbf{a}^t - 2\eta(\mathbf{a}^t \bar{\mathbf{b}}^\top \bar{\mathbf{b}} - \mathbf{a}^* {\bar{\mathbf{b}}^*}^\top \bar{\mathbf{b}})}{\|\hat{\mathbf{a}}^+\|} \tag{204}$$

where $\bar{\mathbf{b}}^* = \sum_{i=1}^N \mathbf{b}_i^*$, $\|\hat{\mathbf{a}}^+\| = \|\mathbf{a}^t - 2\eta(\mathbf{a}^t \bar{\mathbf{b}}^\top \bar{\mathbf{b}} - \mathbf{a}^* {\bar{\mathbf{b}}^*}^\top \bar{\mathbf{b}})\|$.

Proof. **Minimization for** b_i**.** At the start of each iteration, each client computes the analytic solution for $\mathbf{b}_i$ by fixing $\mathbf{a}$ and solving their local objective $\arg\min_{\mathbf{b}_i} \|\mathbf{a}^* {\mathbf{b}_i^*}^\top - \mathbf{a}\mathbf{b}_i^\top\|^2$, where $\mathbf{a}^*$ and $\mathbf{a}$ are both unit vectors. Setting $\mathbf{a} = \mathbf{a}^t$, we obtain $\mathbf{b}_i$ such that

$$\mathbf{b}_i = \frac{\mathbf{b}_i^* {\mathbf{a}^*}^\top \mathbf{a}}{\mathbf{a}^\top \mathbf{a}} = \mathbf{b}_i^* {\mathbf{a}^*}^\top \mathbf{a} \tag{205}$$

(205) follows since $\mathbf{a}^\top \mathbf{a} = 1$.

Aggregation for $\mathbf{b}_i$**.** The server simply computes the average of $\{\mathbf{b}_i\}_{i=1}^N$ and gets

$$\bar{\mathbf{b}} = \sum_{i=1}^N \mathbf{b}_i = \sum_{i=1}^N \mathbf{b}_i^* {\mathbf{a}^*}^\top \mathbf{a} = \bar{\mathbf{b}}^* {\mathbf{a}^*}^\top \mathbf{a} \tag{206}$$

The server then sends $\bar{\mathbf{b}}$ to clients for synchronization.

Gradient Descent for $\hat{\mathbf{a}}$**.** In this step, each client fixes $\mathbf{b}_i$ to $\bar{\mathbf{b}}$ received from the server and update $\mathbf{a}$ using gradient descent. With the following gradient

$$\nabla_{\mathbf{a}} l_i(\mathbf{a}, \bar{\mathbf{b}}) = 2(\mathbf{a}\bar{\mathbf{b}}^\top \bar{\mathbf{b}} - \mathbf{a}^* {\mathbf{b}_i^*}^\top \bar{\mathbf{b}}) \tag{207}$$

Thus, with step size η, $\mathbf{a}$ is updated such as

$$\begin{aligned}
\hat{\mathbf{a}}^+ &= \mathbf{a} - \frac{\eta}{N} \sum_{i=1}^N \nabla_{\mathbf{a}} l_i(\mathbf{a}, \bar{\mathbf{b}}) \\
&= \mathbf{a} - 2\frac{\eta}{N} \sum_{i=1}^N (\mathbf{a}\bar{\mathbf{b}}^\top \bar{\mathbf{b}} - \mathbf{a}^* {\mathbf{b}_i^*}^\top \bar{\mathbf{b}}) \\
&= \mathbf{a} - 2\eta(\mathbf{a}\bar{\mathbf{b}}^\top \bar{\mathbf{b}} - \mathbf{a}^* {\bar{\mathbf{b}}^*}^\top \bar{\mathbf{b}}) \qquad (208) \\
\hat{\mathbf{a}} &= \frac{\mathbf{a} - 2\eta(\mathbf{a}\bar{\mathbf{b}}^\top \bar{\mathbf{b}} - \mathbf{a}^* {\bar{\mathbf{b}}^*}^\top \bar{\mathbf{b}})}{\|\hat{\mathbf{a}}^+\|} \qquad (209)
\end{aligned}$$

□

Lemma A3.16. *Let* $\delta_t = |\sin\theta(\mathbf{a}^*, \mathbf{a}^t)|$ *be the angle distance between* $\mathbf{a}^*$ *and* $\mathbf{a}^t$*. Assume that Assumption A3.12 holds and* $\delta_t \leq \delta_{t-1} \leq \cdots \leq \delta_0$*, if* $\eta \leq \frac{1}{2L_{\max}^2}$*, then*

$$|\sin\theta(\mathbf{a}^{t+1}, \mathbf{a}^*)| = \delta_{t+1} \leq \delta_t \cdot (1 - 2\eta(1 - (\delta^0)^2)\|\bar{\mathbf{b}}^*\|^2) \tag{210}$$

Proof. From Lemma A3.15, $\mathbf{a}^{t+1}$ and $\mathbf{b}^{t+1}$ are computed as follows:

$$\mathbf{b}^{t+1} = \bar{\mathbf{b}} = \bar{\mathbf{b}}^* {\mathbf{a}^*}^\top \mathbf{a}^t \tag{211}$$

$$\mathbf{a}^{t+1} = \frac{\mathbf{a}^t - 2\eta(\mathbf{a}^t \bar{\mathbf{b}}^\top \bar{\mathbf{b}} - \mathbf{a}^* {\bar{\mathbf{b}}^*}^\top \bar{\mathbf{b}})}{\|\mathbf{a}^t - 2\eta(\mathbf{a}^t \bar{\mathbf{b}}^\top \bar{\mathbf{b}} - \mathbf{a}^* {\bar{\mathbf{b}}^*}^\top \bar{\mathbf{b}})\|} \tag{212}$$

Note that $\mathbf{a}^t$ and $\mathbf{a}^{t+1}$ are both unit vectors. Now, we multiply both sides of Equation (212) by the projection operator $\mathbf{P} = \mathbf{I}_d - \mathbf{a}^*(\mathbf{a}^*)^\top$, which is the projection to the direction orthogonal to $\mathbf{a}^*$. We obtain:

$$\begin{aligned}
\mathbf{P}\mathbf{a}^{t+1} &= \frac{\mathbf{P}\mathbf{a}^t - 2\eta \mathbf{P}\mathbf{a}^t \bar{\mathbf{b}}^\top \bar{\mathbf{b}} + \mathbf{P}\mathbf{a}^* {\bar{\mathbf{b}}^*}^\top \bar{\mathbf{b}}}{\|\mathbf{a}^t - 2\eta(\mathbf{a}^t \bar{\mathbf{b}}^\top \bar{\mathbf{b}} - \mathbf{a}^* {\bar{\mathbf{b}}^*}^\top \bar{\mathbf{b}})\|} \qquad (213) \\
&= \frac{\mathbf{P}\mathbf{a}^t - 2\eta \mathbf{P}\mathbf{a}^t \bar{\mathbf{b}}^\top \bar{\mathbf{b}}}{\|\mathbf{a}^t - 2\eta(\mathbf{a}^t \bar{\mathbf{b}}^\top \bar{\mathbf{b}} - \mathbf{a}^* {\bar{\mathbf{b}}^*}^\top \bar{\mathbf{b}})\|} \qquad (214)
\end{aligned}$$

The third term of the numerator is canceled since $\mathbf{P}\mathbf{a}^* = (\mathbf{I}_d - \mathbf{a}^*(\mathbf{a}^*)^\top)\mathbf{a}^* = 0$. Thus,

$$\|\mathbf{P}\mathbf{a}^{t+1}\| \le \frac{\|\mathbf{P}\mathbf{a}^t\||1-2\eta\bar{\mathbf{b}}^\top\bar{\mathbf{b}}|}{\|\mathbf{a}^t - 2\eta(\mathbf{a}^t\bar{\mathbf{b}}^\top\bar{\mathbf{b}} - \mathbf{a}^*\bar{\mathbf{b}}^{*\top}\bar{\mathbf{b}})\|} \tag{215}$$

Let $\delta_t = |\sin\theta(\mathbf{a}^*, \mathbf{a}^t)|$. Equation (213) becomes:

$$\delta_{t+1} \le \delta^t \frac{|1-2\eta\bar{\mathbf{b}}^\top\bar{\mathbf{b}}|}{\|\mathbf{a}^t - 2\eta(\mathbf{a}^t\bar{\mathbf{b}}^\top\bar{\mathbf{b}} - \mathbf{a}^*\bar{\mathbf{b}}^{*\top}\bar{\mathbf{b}})\|} \tag{216}$$

$$= \delta_t \frac{|1-2\eta\bar{\mathbf{b}}^\top\bar{\mathbf{b}}|}{\|\mathbf{a}^t(1-2\eta\bar{\mathbf{b}}^\top\bar{\mathbf{b}}) + 2\eta\mathbf{a}^*\bar{\mathbf{b}}^{*\top}\bar{\mathbf{b}}\|} \tag{217}$$

$$= \delta_t C \tag{218}$$

Obviously $C \ge 0$. We drop the superscript t when it is clear from context. Note that we have

$$C^2 = \frac{|1-2\eta\bar{\mathbf{b}}^\top\bar{\mathbf{b}}|^2}{\|\mathbf{a}(1-2\eta\bar{\mathbf{b}}^\top\bar{\mathbf{b}}) + 2\eta\mathbf{a}^*\bar{\mathbf{b}}^{*\top}\bar{\mathbf{b}}\|^2} \tag{219}$$

$$= \frac{|1-2\eta\bar{\mathbf{b}}^\top\bar{\mathbf{b}}|^2}{(1-2\eta\bar{\mathbf{b}}^\top\bar{\mathbf{b}})^2\mathbf{a}^\top\mathbf{a} + 4\eta^2(\bar{\mathbf{b}}^{*\top}\bar{\mathbf{b}})^2 + 4\eta(1-2\eta\bar{\mathbf{b}}^\top\bar{\mathbf{b}})\mathbf{a}^\top\mathbf{a}^*\bar{\mathbf{b}}^{*\top}\bar{\mathbf{b}}} \tag{220}$$

$$= \frac{|1-2\eta\bar{\mathbf{b}}^\top\bar{\mathbf{b}}|^2}{(1-2\eta\bar{\mathbf{b}}^\top\bar{\mathbf{b}})^2 + 4\eta^2(\bar{\mathbf{b}}^{*\top}\bar{\mathbf{b}})^2 + 4\eta(1-2\eta\bar{\mathbf{b}}^\top\bar{\mathbf{b}})\mathbf{a}^\top\mathbf{a}^*\bar{\mathbf{b}}^{*\top}\bar{\mathbf{b}}} \tag{221}$$

Recall that $\bar{\mathbf{b}} = \bar{\mathbf{b}}^*\mathbf{a}^{*\top}\mathbf{a} = \bar{\mathbf{b}}^*\cos\theta(\mathbf{a}^*, \mathbf{a})$, (221) becomes:

$$C^2 = \frac{|1-2\eta\bar{\mathbf{b}}^\top\bar{\mathbf{b}}|^2}{(1-2\eta\bar{\mathbf{b}}^\top\bar{\mathbf{b}})^2 + 4\eta^2(\bar{\mathbf{b}}^{*\top}\bar{\mathbf{b}})^2 + 4\eta(1-2\eta\bar{\mathbf{b}}^\top\bar{\mathbf{b}})\mathbf{a}^\top\mathbf{a}^*\bar{\mathbf{b}}^{*\top}\bar{\mathbf{b}}} \tag{222}$$

$$= \frac{|1-2\eta\bar{\mathbf{b}}^\top\bar{\mathbf{b}}|^2}{1 + 4\eta^2\bar{\mathbf{b}}^\top\bar{\mathbf{b}}(\bar{\mathbf{b}}^{*\top}\bar{\mathbf{b}}^* - \bar{\mathbf{b}}^\top\bar{\mathbf{b}})} \tag{223}$$

$$\le (1-2\eta\bar{\mathbf{b}}^\top\bar{\mathbf{b}})^2 \tag{224}$$

$$= (1-2\eta\|\bar{\mathbf{b}}\|^2)^2 \tag{225}$$

where (224) holds because $\bar{\mathbf{b}}^{*\top}\bar{\mathbf{b}}^* - \bar{\mathbf{b}}^\top\bar{\mathbf{b}} = (1-\cos^2\theta(\mathbf{a}^*, \mathbf{a}))\bar{\mathbf{b}}^{*\top}\bar{\mathbf{b}}^* \ge 0$. Equation (225) implies $C \le 1 - 2\eta\|\bar{\mathbf{b}}\|^2$ if $2\eta\|\bar{\mathbf{b}}\|^2 \le 1$, which can be ensured by choosing a proper step size $\eta \le \frac{1}{2L_{max}^2} \le \frac{1}{2\|\bar{\mathbf{b}}\|^2}$. Now by the assumption that $\delta_t \le \delta_{t-1} \le \cdots \le \delta_0$,

$$C \le 1 - 2\eta\|\bar{\mathbf{b}}\|^2 \tag{226}$$

$$= 1 - 2\eta\cos^2\theta(\mathbf{a}^*, \mathbf{a})\|\bar{\mathbf{b}}^*\|^2 \tag{227}$$

$$= 1 - 2\eta(1-(\delta^t)^2)\|\bar{\mathbf{b}}^*\|^2 \tag{228}$$

$$\le 1 - 2\eta(1-(\delta^0)^2)\|\bar{\mathbf{b}}^*\|^2 \tag{229}$$

Summarizing, we obtain $\delta^{t+1} \le \delta^t C \le \delta^t(1-2\eta(1-(\delta^0)^2)\|\bar{\mathbf{b}}^*\|^2)$. □

Proposition A3.17. *(FFA-LoRA lower bound) Suppose we are in the setting described in Section A3.5. For any set of ground truth parameters ($\mathbf{a}^*, \{\mathbf{b}_i^*\}_{i=1}^N$), initialization $\mathbf{a}^0$, initial angle distance $\delta_0 \in (0,1)$, we apply Freezing-A scheme to obtain a shared global model ($\mathbf{a}^0, \mathbf{b}^{FFA}$), where $\mathbf{b}^{FFA} = \mathbf{b}^*\mathbf{a}^{*\top}\mathbf{a}^0$. The global loss is*

$$\frac{1}{N}\sum_{i=1}^{N} l_i(\mathbf{a}^0, \mathbf{b}^{FFA}) = \gamma^2 + \|\bar{\mathbf{b}}^*\|^2\delta_0^2 \tag{230}$$

Proof. Through single step of minimization on $\mathbf{b}_i$ and corresponding aggregation, the minimum of the global objective is reached by FFA-LoRA. $\mathbf{b}^{FFA}$ is obtained through:

$$\mathbf{b}_i = \frac{\mathbf{b}_i^*\mathbf{a}^{*\top}\mathbf{a}^0}{\mathbf{a}^{0\top}\mathbf{a}^0} = \mathbf{b}_i^*\mathbf{a}^{*\top}\mathbf{a}^0 \tag{231}$$

$$\mathbf{b}^{FFA} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{b}_i = \bar{\mathbf{b}}^*\mathbf{a}^{*\top}\mathbf{a}^0 \tag{232}$$

Next we compute the global loss with a shared global model $(\mathbf{a}^0, \bar{\mathbf{b}}^{FFA})$. Note that we use Tr(.) to denote the trace of a matrix.

$$\frac{1}{N}\sum_{i=1}^{N} l_i(\mathbf{a}^0, \mathbf{b}^{FFA}) \tag{233}$$
$$= \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{a}^*(\mathbf{b}_i^*)^\top - \mathbf{a}^0(\mathbf{b}^{FFA})^\top\|^2 \tag{234}$$
$$= \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{a}^*(\mathbf{b}_i^*)^\top - \mathbf{a}^*(\bar{\mathbf{b}}^*)^\top + \mathbf{a}^*(\bar{\mathbf{b}}^*)^\top - \mathbf{a}^0(\mathbf{b}^{FFA})^\top\|^2 \tag{235}$$
$$= \frac{1}{N}\sum_{i=1}^{N} (\|\mathbf{a}^*(\mathbf{b}_i^*)^\top - \mathbf{a}^*(\bar{\mathbf{b}}^*)^\top\|^2 + \|\mathbf{a}^*(\bar{\mathbf{b}}^*)^\top - \mathbf{a}^0(\mathbf{b}^{FFA})^\top\|^2$$
$$+ 2\mathrm{Tr}((\mathbf{a}^*(\mathbf{b}_i^*)^\top - \mathbf{a}^*(\bar{\mathbf{b}}^*)^\top)^\top(\mathbf{a}^*(\bar{\mathbf{b}}^*)^\top - \mathbf{a}^0(\mathbf{b}^{FFA})^\top)) \tag{236}$$
$$= \frac{1}{N}\sum_{i=1}^{N} (\|\mathbf{a}^*(\mathbf{b}_i^*)^\top - \mathbf{a}^*(\bar{\mathbf{b}}^*)^\top\|^2 + \|\mathbf{a}^*(\bar{\mathbf{b}}^*)^\top - \mathbf{a}^0(\mathbf{b}^{FFA})^\top\|^2)$$
$$+ 2\mathrm{Tr}((\mathbf{a}^*\frac{1}{N}\sum_{i=1}^{N}(\mathbf{b}_i^*)^\top - \mathbf{a}^*(\bar{\mathbf{b}}^*)^\top)^\top(\mathbf{a}^*(\bar{\mathbf{b}}^*)^\top - \mathbf{a}^0(\mathbf{b}^{FFA})^\top)) \tag{237}$$
$$= \frac{1}{N}\sum_{i=1}^{N} (\|\mathbf{a}(\mathbf{b}_i^* - \bar{\mathbf{b}}^*)^\top\|^2 + \|\mathbf{a}^*(\bar{\mathbf{b}}^*)^\top - \mathbf{a}^0{\mathbf{a}^0}^\top\mathbf{a}^*(\bar{\mathbf{b}}^*)^\top\|^2) \tag{238}$$
$$= \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{b}_i^* - \bar{\mathbf{b}}^*\|^2 + \frac{1}{N}\sum_{i=1}^{N} \|(\mathbf{I}_d - \mathbf{a}^0{\mathbf{a}^0}^\top)\mathbf{a}^*(\bar{\mathbf{b}}^*)^\top\|^2 \tag{239}$$
$$= \frac{1}{N}\sum_{i=1}^{N} \|\mathbf{b}_i^* - \bar{\mathbf{b}}^*\|^2 + \frac{1}{N}\sum_{i=1}^{N} \|(\mathbf{I}_d - \mathbf{a}^0{\mathbf{a}^0}^\top)\mathbf{a}^*\|^2\|\bar{\mathbf{b}}^*\|^2 \tag{240}$$
$$= \gamma^2 + \|\bar{\mathbf{b}}^*\|^2\delta_0^2 \tag{241}$$

where (238) holds since the last term is 0, (239) and (240) hold since $\|\mathbf{u}\mathbf{v}^\top\| = \|\mathbf{u}\| \cdot \|\mathbf{v}^\top\|$ for vector $\mathbf{u}$ and $\mathbf{v}$, (241) holds because of Definition A3.13. □

Proof of Theorem A.14

Proof. In order to achieve ϵ-recovery of $\mathbf{a}^*$, we need

$$\delta^0(1 - c(1 - (\delta^0)^2))^T \leq \epsilon \tag{242}$$
$$(1 - c(1 - (\delta^0)^2))^T \leq \frac{\epsilon}{\delta^0} \tag{243}$$
$$T \log(1 - c(1 - (\delta^0)^2)) \leq \log(\frac{\epsilon}{\delta^0}) \tag{244}$$
$$\tag{245}$$

We proceed such that

$$T \geq \frac{\log(\frac{\epsilon}{\delta^0})}{\log(1 - c(1 - (\delta^0)^2))} \tag{246}$$
$$> \frac{\log(\frac{\epsilon}{\delta^0})}{-c(1 - (\delta^0)^2)} \tag{247}$$
$$= \frac{1}{c(1 - (\delta^0)^2)} \log(\frac{\delta^0}{\epsilon}) \tag{248}$$

where (247) follows by using $\log(1 - x) < -x$ for $|x| < 1$.

Now we show the convergence to the global minimizer. Recall that $\mathbf{b}^{T+1} = \bar{\mathbf{b}}^* {\mathbf{a}^*}^\top \mathbf{a}^T$ and $\delta^T = \|(\mathbf{I}_d - \mathbf{a}^T(\mathbf{a}^T)^\top)\mathbf{a}^*\|$, we have

$$
\begin{aligned}
\|\mathbf{a}^T(\mathbf{b}^{T+1})^\top - \mathbf{a}^*{\bar{\mathbf{b}}^*}^\top\| &= \|\mathbf{a}^T(\mathbf{a}^T)^\top\mathbf{a}^*{\bar{\mathbf{b}}^*}^\top - \mathbf{a}^*{\bar{\mathbf{b}}^*}^\top\| && (249)\\
&= \|(\mathbf{a}^T(\mathbf{a}^T)^\top - \mathbf{I}_d)\mathbf{a}^*{\bar{\mathbf{b}}^*}^\top\| && (250)\\
&= \|(\mathbf{I}_d - \mathbf{a}^T(\mathbf{a}^T)^\top)\mathbf{a}^*\| \cdot \|\bar{\mathbf{b}}^*\| && (251)\\
&\leq \epsilon\|\bar{\mathbf{b}}^*\| && (252)\\
&= \epsilon\|\mathbf{a}^*{\bar{\mathbf{b}}^*}^\top\| && (253)
\end{aligned}
$$

where (253) is due to the fact that $\|\mathbf{x}\mathbf{y}^\top\| = \|\mathbf{x}\|\|\mathbf{y}\|$ and $\|\mathbf{a}^*\| = 1$.

□

Proposition A3.17 shows that for any $\delta_0 \in (0, 1)$, the global objective of FFA-LoRA is given by (241), comprising two terms: γ^2, reflecting the heterogeneity of $\{\mathbf{b}_i^*\}_{i=1}^N$, and $\|\bar{\mathbf{b}}^*\|^2\delta_0^2$, due to the angular distance between $\mathbf{a}^0$ and $\mathbf{a}^*$. By Theorem A3.14, RoLoRA achieves ϵ-accurate recovery of the global minimizer, with global loss upper bounded by $\gamma^2 + \|\bar{\mathbf{b}}^*\|^2\epsilon^2$, since RoLoRA reduces the angular distance loss from $\|\bar{\mathbf{b}}^*\|^2\delta_0^2$ to $\|\bar{\mathbf{b}}^*\|^2\epsilon^2$. We can make ϵ arbitrarily small by increasing the iterations.

A4 Convergence Analysis of Non-Convex Case

We follow the approach of Li et al. [25] to demonstrate the convergence of RoLoRA (Algorithm 1) in smooth, non-convex landscapes. Assumptions A4.1 and A4.2 are standard and commonly employed in the convergence analysis of federated learning. Assumption A4.3 is adapted from FedSA-LoRA [15], which proposes a personalized federated fine-tuning framework that maintains local diversity in $\mathbf{B}$ while aggregating only $\mathbf{A}$ through simultaneous updates of both $\mathbf{A}$ and $\mathbf{B}$. In contrast, our work focuses on a single global model with alternating optimization of $\mathbf{A}$ and $\mathbf{B}$.

In Assumption A4.3, bounding Frobenius norms of $\mathbf{A}$ and $\mathbf{B}$ is a standard weight-regularity requirement in LoRA fine-tuning practice, where small rank and scaling factors keep the adapters from exploding. The inner-product conditions simply posit that each adapter has enough non-degenerate singular values, requiring that the low-rank updates retain sufficient rank and alignment with local gradient directions. Formally, we require the smallest singular values of $\mathbf{A}_i^t$ and $\mathbf{B}_i^t$ to be lower bounded by $\sqrt{c_A}$ and $\sqrt{c_B}$.

Assumption A4.1 (Bounded Stochastic Gradient)**.** Let a mini-batch $\mathbf{x}_i$, be drawn uniformly at random from client i's dataset, meaning $\mathbb{E}_{\mathbf{x}_i}[\nabla_{\mathbf{W}_i} l_i(\mathbf{W}_i, \mathbf{x}_i)] = \nabla_{\mathbf{W}_i} l_i(\mathbf{W}_i)$. We assume that the expected squared norm of any stochastic gradient is uniformly bounded, that is,

$$\mathbb{E}_{\mathbf{x}_i}\left\|\nabla_{\mathbf{W}} l_i(\mathbf{W}, \mathbf{x}_i)\right\|^2 \leq G^2$$

where the expectation is over the random draw of the mini-batch $\mathbf{x}_i$, and $G > 0$ is a constant.

Assumption A4.2 (Lipschitz smooth)**.** Loss functions $l_1, \cdots, l_N$ are all L-smooth. For all weights $\mathbf{W}$ and $\mathbf{U}$:

$$l_i(\mathbf{V}) \leq l_i(\mathbf{W}) + \langle \nabla_{\mathbf{W}} l_i(\mathbf{W}), \mathbf{V} - \mathbf{W}\rangle_F + \frac{L}{2}\|\mathbf{V} - \mathbf{W}\|_F^2, \forall i \in [N]$$

Assumption A4.3. Let $\mathbf{W}_i = \mathbf{W}_0 + \mathbf{B}_i\mathbf{A}_i$ represent the model parameters for the i-th client. There exist constants $C_B > 0$, $C_A > 0$, $c_B > 0$, and $c_A > 0$ such that:

$$\begin{aligned}
\|\mathbf{B}_i\|_F &\leq C_B, \\
\|\mathbf{A}_i\|_F &\leq C_A, \\
\langle \mathbf{A}_i{\mathbf{A}_i}^\top, \nabla_{\mathbf{W}} l_i(\mathbf{W}_i)\nabla_{\mathbf{W}} l_i(\mathbf{W}_i)^\top\rangle_F &\geq c_A\|\nabla_{\mathbf{W}} l_i(\mathbf{W}_i)\|_F^2, \\
\langle {\mathbf{B}_i}^\top\mathbf{B}_i, \nabla_{\mathbf{W}} l_i(\mathbf{W}_i)^\top\nabla_{\mathbf{W}} l_i(\mathbf{W}_i)\rangle_F &\geq c_B\|\nabla_{\mathbf{W}} l_i(\mathbf{W}_i)\|_F^2,
\end{aligned}$$

for all $i \in [N]$.

Theorem A4.4 (Convergence to the stationary point)**.** *Let Assumption A4.1, A4.2, and A4.3 hold. Suppose each client runs $2T$ rounds, each consisting of Q local epochs, using a learning rate $\eta \propto O(1/\sqrt{T})$, then we obtain:*

$$\min_{0\leq t\leq 2T} \mathbb{E}[\|\nabla_{\mathbf{W}} l_i(\mathbf{W}^t)\|_F^2] \leq \frac{\Delta_i}{2T\eta c_{min}} + \frac{D\eta}{2c_{min}} \tag{254}$$

where $\Delta_i = \mathbb{E}[l_i(\mathbf{W}^0)] - l_i^*, c_{min} = \min(c_A, c_B)$, D is chosen such that $D\eta^2 \geq D_A + D_B$, and $D_A = L\eta C_A^2Q^2G^2 + \frac{1}{2}\eta G^2 + 2L\eta^2C_A^2Q^2G^2 + \frac{L}{2}\eta^2C_A^4G^2, D_B = L\eta C_B^2Q^2G^2 + \frac{1}{2}\eta G^2 + 2L\eta^2C_B^2Q^2G^2 + \frac{L}{2}\eta^2C_B^4G^2$.

According to Theorem A4.4, we achieve an $O\left(\frac{1}{\sqrt{T}}\right)$ convergence rate toward a stationary point under smooth, non-convex conditions, matching the convergence rate of FedAVG in the same setting. We follow a similar derivation framework to [15], adopting similar proof techniques, such as applying Assumption A4.2 to the global and local models (Eq.(262) and Eq.(276)). However, due to structural differences between our algorithm and that of [15]—in particular, our use of alternating optimization rather than simultaneous updates—the derivation diverges in how key gradient bounds are applied and combined. As a result, even though the proof steps are analogous, our final convergence bound depends on $\min(c_A, c_B)$ and features decoupled $\mathbf{A}$- and $\mathbf{B}$-related terms, in contrast to the $c_A + c_B$ dependence along with cross-terms due to the simultaneous updates of $\mathbf{A}$ and $\mathbf{B}$ in [15]. Although FedSA-LoRA reports the same $O\left(\frac{1}{\sqrt{T}}\right)$ convergence rate, the two algorithms address fundamentally different FL scenarios: FedSA-LoRA optimizes personalized client models, whereas RoLoRA learns a single shared global model.

Proof. Let $\mathbf{W}^{2t} = \mathbf{W}_0 + \mathbf{A}^t\mathbf{B}^t$ be the global model parameters at the $2t$-th communication round. Let $\mathbf{W}_{i,q}^{2t} = \mathbf{W}_0 + \mathbf{A}^t\mathbf{B}_{i,q}^t$ be the local model parameters of client i at the q-th local epoch of the $2t$-th communication round, where each client performs a total of Q local epochs. We define the following for convenience:

The $2t$-th communication round:

$$\mathbf{W}^{2t} = \mathbf{W}_0 + \mathbf{A}^t\mathbf{B}^t \tag{255}$$

The q-th local epoch of the $2t$-th communication round:

$$\mathbf{W}_{i,q}^{2t} = \mathbf{W}_0 + \mathbf{A}^t\mathbf{B}_{i,q}^t \tag{256}$$

The $2t+1$-th communication round:

$$\mathbf{W}^{2t+1} = \mathbf{W}_0 + \mathbf{A}^t\mathbf{B}^{t+1} = \mathbf{W}_0 + \frac{1}{N}\sum_{i=1}^{N}\mathbf{A}^t\mathbf{B}_{i,Q}^t \tag{257}$$

The q-th local epoch of the $2t+1$-th communication round:

$$\mathbf{W}_{i,q}^{2t+1} = \mathbf{W}_0 + \mathbf{A}_{i,q}^t\mathbf{B}^{t+1} \tag{258}$$

The $2t+2$-th communication round:

$$\mathbf{W}^{2t+2} = \mathbf{W}_0 + \mathbf{A}^{t+1}\mathbf{B}^{t+1} = \mathbf{W}_0 + \frac{1}{N}\sum_{i=1}^{N}\mathbf{A}_{i,Q}^t\mathbf{B}^{t+1} \tag{259}$$

According to chain rule,

$$\nabla_{\mathbf{B}} l_i(\mathbf{W}) = \mathbf{A}^\top \nabla_{\mathbf{W}} l_i(\mathbf{W}) \tag{260}$$

$$\nabla_{\mathbf{A}} l_i(\mathbf{W}) = \nabla_{\mathbf{W}} l_i(\mathbf{W})\mathbf{B}^\top \tag{261}$$

Now we apply Assumption A4.2 to local update of $\mathbf{B}$, and get

$$l_i(\mathbf{W}_{i,1}^{2t}) \le l_i(\mathbf{W}^{2t}) + \langle \mathbf{W}_{i,1}^{2t} - \mathbf{W}^{2t}, \nabla_{\mathbf{W}} l_i(\mathbf{W}^{2t})\rangle_F + \frac{L}{2}\|\mathbf{W}_{i,1}^{2t} - \mathbf{W}^{2t}\|_F^2 \tag{262}$$

By (255) and (256),

$$\begin{aligned}\mathbf{W}_{i,1}^{2t} - \mathbf{W}^{2t} &= \mathbf{A}^t(\mathbf{B}_{i,1}^t - \mathbf{B}^t) && (263)\\ &= -\eta\mathbf{A}^t{\mathbf{A}^t}^\top \nabla_{\mathbf{W}} l_i(\mathbf{W}^{2t}, \mathbf{x}_{i,1}^{2t}) && (264)\end{aligned}$$

where η is the learning rate. Then

$$\begin{aligned}l_i(\mathbf{W}_{i,1}^{2t}) \le l_i(\mathbf{W}^{2t}) &- \eta\langle \mathbf{A}^t{\mathbf{A}^t}^\top \nabla_{\mathbf{W}} l_i(\mathbf{W}^{2t}, \mathbf{x}_{i,1}^{2t}), \nabla_{\mathbf{W}} l_i(\mathbf{W}^{2t})\rangle_F \\ &+ \frac{L}{2}\eta^2\|\mathbf{A}^t{\mathbf{A}^t}^\top \nabla_{\mathbf{W}} l_i(\mathbf{W}^{2t}, \mathbf{x}_{i,1}^{2t})\|_F^2 \end{aligned} \tag{265}$$

Taking expectation of (265),

$$\mathbb{E}[l_i(\mathbf{W}_{i,1}^{2t})] \le \mathbb{E}[l_i(\mathbf{W}^{2t})] - \eta\langle \mathbf{A}^t{\mathbf{A}^t}^\top \nabla_{\mathbf{W}} l_i(\mathbf{W}^{2t}), \nabla_{\mathbf{W}} l_i(\mathbf{W}^{2t})\rangle_F + \frac{L}{2}\eta^2\mathbb{E}[\|\mathbf{A}^t{\mathbf{A}^t}^\top \nabla_{\mathbf{W}} l_i(\mathbf{W}^{2t}, \mathbf{x}_{i,1}^{2t})\|_F^2] \tag{266}$$

The inner product term is lower bounded such that

$$\begin{aligned}\langle \mathbf{A}^t{\mathbf{A}^t}^\top \nabla_{\mathbf{W}} l_i(\mathbf{W}^{2t}), \nabla_{\mathbf{W}} l_i(\mathbf{W}^{2t})\rangle_F &= \mathrm{Tr}[\nabla_{\mathbf{W}} l_i(\mathbf{W}^{2t})^\top \mathbf{A}^t{\mathbf{A}^t}^\top \nabla_{\mathbf{W}} l_i(\mathbf{W}^{2t})] && (267)\\ &= \mathrm{Tr}[\mathbf{A}^t{\mathbf{A}^t}^\top \nabla_{\mathbf{W}} l_i(\mathbf{W}^{2t})\nabla_{\mathbf{W}} l_i(\mathbf{W}^{2t})^\top] && (268)\\ &= \langle \mathbf{A}^t{\mathbf{A}^t}^\top, \nabla_{\mathbf{W}} l_i(\mathbf{W}^{2t})\nabla_{\mathbf{W}} l_i(\mathbf{W}^{2t})^\top\rangle_F && (269)\\ &\ge c_A\|\nabla_{\mathbf{W}} l_i(\mathbf{W}^{2t})\|_F^2 && (270)\end{aligned}$$

where (270) follows by Assumption A4.3. Moreover,

$$\frac{L}{2}\eta^2\mathbb{E}[\|\mathbf{A}^t{\mathbf{A}^t}^\top\nabla_{\mathbf{W}}l_i(\mathbf{W}^{2t},\mathbf{x}_{i,1}^{2t})\|_F^2] \leq \frac{L}{2}\eta^2\mathbb{E}[\|\mathbf{A}^t\|_F^4\cdot\|\nabla_{\mathbf{W}}l_i(\mathbf{W}^{2t},\mathbf{x}_{i,1}^{2t})\|_F^2] \tag{271}$$

$$\leq \frac{L}{2}\eta^2\mathbb{E}[C_A^4\cdot\|\nabla_{\mathbf{W}}l_i(\mathbf{W}^{2t},\mathbf{x}_{i,1}^{2t})\|_F^2] \tag{272}$$

$$= \frac{L}{2}\eta^2C_A^4\mathbb{E}[\|\nabla_{\mathbf{W}}l_i(\mathbf{W}^{2t},\mathbf{x}_{i,1}^{2t})\|_F^2] \tag{273}$$

$$\leq \frac{L}{2}\eta^2C_A^4G^2 \tag{274}$$

where (271) follows by Assumption A4.3, (274) follows by Assumption A4.1. Combining (266), (270), and (274), we get

$$\mathbb{E}[l_i(\mathbf{W}_{i,1}^{2t})] \leq \mathbb{E}[l_i(\mathbf{W}^{2t})] - \eta c_A\|\nabla_{\mathbf{W}}l_i(\mathbf{W}^{2t})\|_F^2 + \frac{L}{2}\eta^2C_A^4G^2 \tag{275}$$

Next we apply Assumption A4.3 to aggregation step and take expectation on both sides,

$$\mathbb{E}[l_i(\mathbf{W}^{2t+1})] \leq \mathbb{E}[l_i(\mathbf{W}_{i,1}^{2t})] + \mathbb{E}[\langle\mathbf{W}^{2t+1}-\mathbf{W}_{i,1}^{2t},\nabla_{\mathbf{W}}l_i(\mathbf{W}_{i,1}^{2t})\rangle_F] + \frac{L}{2}\mathbb{E}[\|\mathbf{W}^{2t+1}-\mathbf{W}_{i,1}^{2t}\|_F^2] \tag{276}$$

where

$$\mathbf{W}^{2t+1}-\mathbf{W}_{i,1}^{2t} = \frac{1}{N}\sum_{j=1}^{N}\mathbf{A}^t\mathbf{B}_{j,Q}^t - \mathbf{A}^t\mathbf{B}_{i,1}^t \tag{277}$$

$$= \frac{1}{N}\sum_{j=1}^{N}\mathbf{A}^t(\mathbf{B}_{j,Q}^t - \mathbf{B}_{i,1}^t) \tag{278}$$

We have

$$\mathbf{B}_{j,Q}^t = \mathbf{B}^t - \eta\sum_{q=0}^{Q-1}\nabla_{\mathbf{B}}l_j(\mathbf{W}_{j,q}^{2t},\mathbf{x}_{j,q}^{2t}) \tag{279}$$

$$\mathbf{B}_{i,1}^t = \mathbf{B}^t - \eta\nabla_{\mathbf{B}}l_i(\mathbf{W}^{2t},\mathbf{x}_{i,1}^{2t}) \tag{280}$$

$$\mathbf{B}_{j,Q}^t - \mathbf{B}_{i,1}^t = \eta\sum_{q=0}^{Q-1}(\nabla_{\mathbf{B}}l_i(\mathbf{W}^{2t},\mathbf{x}_{i,1}^{2t}) - \nabla_{\mathbf{B}}l_j(\mathbf{W}_{j,q}^{2t},\mathbf{x}_{j,q}^{2t})) \tag{281}$$

$$= \eta{\mathbf{A}^t}^\top\sum_{q=0}^{Q-1}(\nabla_{\mathbf{W}}l_i(\mathbf{W}^{2t},\mathbf{x}_{i,1}^{2t}) - \nabla_{\mathbf{W}}l_j(\mathbf{W}_{j,q}^{2t},\mathbf{x}_{j,q}^{2t})) \tag{282}$$

Thus,

$$\mathbf{W}^{2t+1}-\mathbf{W}_{i,1}^{2t} = \frac{\eta}{N}\mathbf{A}^t\sum_{j=1}^{N}\sum_{q=0}^{Q-1}(\nabla_{\mathbf{W}}l_i(\mathbf{W}^{2t},\mathbf{x}_{i,1}^{2t}) - \nabla_{\mathbf{W}}l_j(\mathbf{W}_{j,q}^{2t},\mathbf{x}_{j,q}^{2t})) \tag{283}$$

Therefore,

$$\left\|\mathbf{W}^{2t+1}-\mathbf{W}_{i,1}^{2t}\right\|_F^2 = \left\|\frac{\eta}{N}\,\mathbf{A}^t\sum_{j=1}^{N}\sum_{q=0}^{Q-1}\left(\nabla_{\mathbf{W}}l_i(\mathbf{W}^{2t},\mathbf{x}_{i,1}^{2t}) - \nabla_{\mathbf{W}}l_j(\mathbf{W}_{j,q}^{2t},\mathbf{x}_{j,q}^{2t})\right)\right\|_F^2 \tag{284}$$

$$\leq \frac{\eta^2}{N^2}\|\mathbf{A}^t\|_F^2\left\|\sum_{j=1}^{N}\sum_{q=0}^{Q-1}\left(\nabla_{\mathbf{W}}l_i(\mathbf{W}^{2t},\mathbf{x}_{i,1}^{2t}) - \nabla_{\mathbf{W}}l_j(\mathbf{W}_{j,q}^{2t},\mathbf{x}_{j,q}^{2t})\right)\right\|_F^2 \tag{285}$$

$$\leq \frac{\eta^2}{N}C_A^2\sum_{j=1}^{N}\left\|\sum_{q=0}^{Q-1}\left(\nabla_{\mathbf{W}}l_i(\mathbf{W}^{2t},\mathbf{x}_{i,1}^{2t}) - \nabla_{\mathbf{W}}l_j(\mathbf{W}_{j,q}^{2t},\mathbf{x}_{j,q}^{2t})\right)\right\|_F^2 \tag{286}$$

$$\leq \frac{\eta^2}{N}C_A^2Q\sum_{j=1}^{N}\sum_{q=0}^{Q-1}\left\|\nabla_{\mathbf{W}}l_i(\mathbf{W}^{2t},\mathbf{x}_{i,1}^{2t}) - \nabla_{\mathbf{W}}l_j(\mathbf{W}_{j,q}^{2t},\mathbf{x}_{j,q}^{2t})\right\|_F^2 \tag{287}$$

Taking expectation,

$$\mathbb{E}\left[\left\|\mathbf{W}^{2t+1}-\mathbf{W}_{i,1}^{2t}\right\|_F^2\right] \leq \frac{\eta^2}{N}C_A^2 Q\sum_{j=1}^{N}\sum_{q=0}^{Q-1}\mathbb{E}\left[\left\|\nabla_{\mathbf{W}} l_i(\mathbf{W}^{2t},\mathbf{x}_{i,1}^{2t})-\nabla_{\mathbf{W}} l_j(\mathbf{W}_{j,q}^{2t},\mathbf{x}_{j,q}^{2t}))\right\|_F^2\right] \tag{288}$$

For any matrices $\mathbf{A}$ and $\mathbf{B}$ (or vectors), we have

$\|\mathbf{A}-\mathbf{B}\|_F^2 = \|\mathbf{A}\|_F^2+\|\mathbf{B}\|_F^2-2\langle\mathbf{A},\mathbf{B}\rangle_F \;\leq\; \|\mathbf{A}\|_F^2+\|\mathbf{B}\|_F^2+2\|\mathbf{A}\|_F\,\|\mathbf{B}\|_F \;\leq\; 2\|\mathbf{A}\|_F^2+2\|\mathbf{B}\|_F^2.$

Thus,

$$\begin{aligned}
&\mathbb{E}\left[\left\|\nabla_{\mathbf{W}} l_i(\mathbf{W}^{2t},\mathbf{x}_{i,1}^{2t})-\nabla_{\mathbf{W}} l_j(\mathbf{W}_{j,q}^{2t},\mathbf{x}_{j,q}^{2t})\right\|_F^2\right]\\
&\leq 2\mathbb{E}\left[\|\nabla_{\mathbf{W}} l_i(\mathbf{W}^{2t},\mathbf{x}_{i,1}^{2t})\|_F^2\right]+2\mathbb{E}\left[\|\nabla_{\mathbf{W}} l_j(\mathbf{W}_{j,q}^{2t},\mathbf{x}_{j,q}^{2t})\|_F^2\right] &(289)\\
&\leq 4G^2 &(290)
\end{aligned}$$

Leading to

$$\frac{L}{2}\mathbb{E}\left[\left\|\mathbf{W}^{2t+1}-\mathbf{W}_{i,1}^{2t}\right\|_F^2\right] \leq 2L\eta^2 C_A^2 Q^2 G^2 \tag{291}$$

For the inner product term of (276), we have

$$\mathbb{E}\left[\langle\mathbf{W}^{2t+1}-\mathbf{W}_{i,1}^{2t},\nabla_{\mathbf{W}} l_i(\mathbf{W}_{i,1}^{2t})\rangle_F\right] \leq \frac{1}{2\eta}\mathbb{E}\left[\left\|\mathbf{W}^{2t+1}-\mathbf{W}_{i,1}^{2t}\right\|_F^2\right]+\frac{1}{2}\eta\mathbb{E}\left[\left\|\nabla_{\mathbf{W}} l_i(\mathbf{W}_{i,1}^{2t})\right\|_F^2\right] \tag{292}$$

$$\leq L\eta C_A^2 Q^2 G^2+\frac{1}{2}\eta G^2 \tag{293}$$

Combining (276), (291), and (293), we obtain

$$\mathbb{E}[l_i(\mathbf{W}^{2t+1})] \leq \mathbb{E}[l_i(\mathbf{W}_{i,1}^{2t})]+L\eta C_A^2 Q^2 G^2+\frac{1}{2}\eta G^2+2L\eta^2 C_A^2 Q^2 G^2 \tag{294}$$

Combining (275) and (294), we derive

$$\begin{aligned}
&\eta c_A\|\nabla_{\mathbf{W}} l_i(\mathbf{W}^{2t})\|_F^2\\
&\leq \mathbb{E}[l_i(\mathbf{W}^{2t})]-\mathbb{E}[l_i(\mathbf{W}^{2t+1})]+L\eta C_A^2 Q^2 G^2+\frac{1}{2}\eta G^2+2L\eta^2 C_A^2 Q^2 G^2+\frac{L}{2}\eta^2 C_A^4 G^2 &(295)
\end{aligned}$$

Analogously, applying the same analysis to the $(2t+1)$-th communication round, which fixes $\mathbf{B}$ and updates $\mathbf{A}$, introduces key modifications to steps such as Eq.(263) and Eq.(277), which govern the weight updates. These changes propagate through the subsequent steps that depend on the updated weights. As a result, we obtain

$$\begin{aligned}
&\eta c_B\|\nabla_{\mathbf{W}} l_i(\mathbf{W}^{2t+1})\|_F^2\\
&\leq \mathbb{E}[l_i(\mathbf{W}^{2t})]-\mathbb{E}[l_i(\mathbf{W}^{2t+1})]+L\eta C_B^2 Q^2 G^2+\frac{1}{2}\eta G^2+2L\eta^2 C_B^2 Q^2 G^2+\frac{L}{2}\eta^2 C_B^4 G^2 &(296)
\end{aligned}$$

Add the two inequalities and then sum over $t=0,1,\ldots,T-1$, we get

$$\sum_{t=0}^{T-1}\eta(c_A\|\nabla_{\mathbf{W}} l_i(\mathbf{W}^{2t})\|_F^2+c_B\|\nabla_{\mathbf{W}} l_i(\mathbf{W}^{2t+1})\|_F^2) \leq \mathbb{E}[l_i(\mathbf{W}^0)]-\mathbb{E}[l_i(\mathbf{W}^{2T})]+T(D_A+D_B) \tag{297}$$

where

$$D_A = L\eta C_A^2 Q^2 G^2+\frac{1}{2}\eta G^2+2L\eta^2 C_A^2 Q^2 G^2+\frac{L}{2}\eta^2 C_A^4 G^2 \tag{298}$$

$$D_B = L\eta C_B^2 Q^2 G^2+\frac{1}{2}\eta G^2+2L\eta^2 C_B^2 Q^2 G^2+\frac{L}{2}\eta^2 C_B^4 G^2 \tag{299}$$

Assume the per-client loss is bounded below by l_i^*, let

$$\Delta_i = \mathbb{E}[l_i(\mathbf{W}^0)]-l_i^*, \quad c_{min} = \min(c_A,c_B), \tag{300}$$

Choosing D such that $D\eta^2 \geq D_A + D_B$, then

$$\min_{0 \leq t \leq 2T} \mathbb{E}[\|\nabla_{\mathbf{W}} l_i(\mathbf{W}^t)\|_F^2] \leq \frac{\Delta_i}{2T\eta c_{min}} + \frac{D\eta}{2c_{min}} \tag{301}$$

We choose $\eta \propto O(1/\sqrt{T})$ so that the overall convergence rate with a diminishing step size is $O(1/\sqrt{T})$ which matches the canonical convergence speed of stochastic gradient methods in non-convex settings. $\square$

A5 Experiments

A5.1 Impact of Non-Linearity on RoLoRA

Across both linear and non-linear settings, all methods perform similarly, with RoLoRA showing modest improvement in the non-linear case, likely due to its better utilization of the added expressiveness from ReLU.

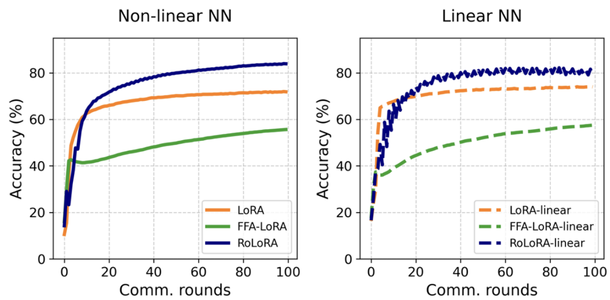

Figure 7: Comparison of RoLoRA, LoRA, and FFA-LoRA on linear and non-linear networks. While overall performance is similar, RoLoRA shows modest gains in the non-linear setting, likely benefiting from ReLU's added expressiveness.

A5.2 Hyper-parameters for GLUE task

	SST-2	QNLI	MNLI	QQP	RTE
Total comm. rounds	500	500	500	500	200
Batch Size	64	32	32	32	32
Local Epochs	20	20	20	20	20

Table 6: Hyper-parameters configurations. Note the total communication rounds are for the setting with 3 clients. When increasing the number of clients, we decrease the total communication rounds accordingly to maintain a constant sample count used during fine-tuning

We show the hyper-parameter configurations for each dataset in Table 6.

A5.3 Effect of Number of Clients

Configurations Table 7 shows the selected layer set attached with LoRA modules for Table 1. We present Table 1 with the results of FlexLoRA [2] added in Table 1.

	Layer Attributes	SST-2	QNLI	MNLI	QQP	RTE
$\mathcal{P}_2$	Type	W_v, W_q	W_v, W_q	W_v, W_q	W_v, W_q	W_v, W_q
	Index	$\{18, \ldots, 23\}$	$\{15, \ldots, 23\}$	$\{15, \ldots, 23\}$	$\{15, \ldots, 23\}$	$\{16, \ldots, 23\}$

Table 7: The selected layer set attached with LoRA modules for Table 1 and Table 2

Rank-2 Results We show the effect of number of clients when using rank-2 LoRA modules in Table 8.

Rank-32 Results In Table 9, we provide additional experiments with rank-32 LoRA adapters in the 20-client and 50-client setting.

Client num	Methods	SST-2	QNLI	MNLI	QQP	RTE
3	LoRA	**95.64**	**92.04**	**85.85**	**86.16**	**82.19**
	FFA-LoRA	94.91	90.11	84.06	85.48	80.86
	RoloRA	95.60	91.62	85.66	**86.16**	**82.19**
20	LoRA	94.27	86.91	81.22	82.07	46.21
	FFA-LoRA	93.92	89.58	80.51	82.62	57.76
	RoloRA	**94.84**	**90.77**	**85.13**	**85.10**	**81.23**
50	LoRA	93.23	82.57	58.96	76.96	49.10
	FFA-LoRA	92.32	85.15	62.79	77.78	53.07
	RoloRA	**94.61**	**89.83**	**85.15**	**85.55**	**72.92**

Table 8: Results with RoBERTa-Large models with varying client numbers (3, 20, 50) using rank-2 LoRA modules in federated setting, maintaining a constant sample count during fine-tuning.

Client num	Methods	MNLI	QQP
20	LoRA	79.72 $\pm$0.38	83.66 $\pm$0.02
	FFA-LoRA	80 $\pm$0.47	84.08 $\pm$0.31
	RoLoRA	85.91 $\pm$0.63	86.37 $\pm$0.09
50	LoRA	70.84 $\pm$4.63	79.75 $\pm$0.31
	FFA-LoRA	74.47 $\pm$1.57	80.65 $\pm$0.31
	RoLoRA	85.46 $\pm$0.08	86.15 $\pm$0.26

Table 9: Results with RoBERTa-Large models in 20-client and 50-client setting using rank-32 LoRA adapters.

10-Client Setting In Table 10, we provide results for 10-client setting with rank-4 adapter. The results show that RoLoRA still outperform other methods. In the 10-client setting, RoLoRA's performance gain over other methods falls between the gains observed in the 3-client and 20-client settings.

	MNLI	QQP	QNLI
LoRA	81.48 $\pm$2.19	84.1 $\pm$0.14	87.73 $\pm$0.67
FFA-LoRA	83.19 $\pm$0.64	84.35 $\pm$0.06	89.88 $\pm$0.13
RoLoRA	84.95 $\pm$0.8	95.25$\pm$0.39	90.3 $\pm$0.76

Table 10: Results with RoBERTa-Large model with 10 clients using rank-4 LoRA adapters, running for 150 rounds in total.

FLoRA vs. RoLoRA Table 11 shows a comparison between FLoRA and RoLoRA. In the 3-client setting, we ran 500 rounds and scaled rounds down proportionally with more clients to keep the total sample budget fixed. RoLoRA consistently outperforms FLoRA across tasks and client counts. While FLoRA eventually converges (e.g., 83.3% on MNLI after 4000 rounds), it does so much more slowly, highlighting RoLoRA's faster convergence and better scalability.

Finetuning Dynamics within 100 Rounds Figure 8 presents the 100-round extension of Figure 3, where RoLoRA consistently converges faster and achieves the highest accuracy.

We want to clarify that Figure 3 focuses on comparing convergence under a fixed sample budget rather than full convergence, and Table 1 shows that this budget suffices for all methods when using 3 clients. However, as shown in Figure 3, with 50 clients, only RoLoRA fully converges, underscoring its efficiency in low-resource settings.

A5.4 Effect of Number of LoRA Parameters

In Table 12, we include the details about layers attached with LoRA adapters for different budget of finetuning parameters, for each dataset.

Client num	Method	MNLI	QQP	QNLI
3	FLoRA	39.29	51.05	59.88
	RoLoRA	**85.70**	**86.14**	**91.64**
20	FLoRA	32.01	51.58	49.89
	RoLoRA	**85.28**	**85.83**	**90.35**
50	FLoRA	31.97	50.54	38.82
	RoLoRA	**82.98**	**85.71**	**90.00**

Table 11: Results with RoBERTa-Large models with rank-4 LoRA adapter for varying numbers of clients (3, 20, 50), comparing FLoRA with RoLoRA, maintaining a constant sample count during finetuning. In the 3-client setting, while FLoRA eventually converges (e.g., 83.3% on MNLI after 4000 rounds), the figure shows results for only 500 rounds, within which FLoRA has not yet converged. This highlights RoLoRA's faster convergence and better scalability.

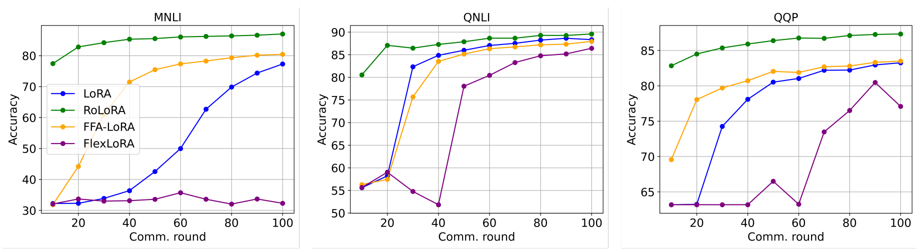

Figure 8: Accuracies over 100 rounds. It involves 50 clients using rank 4.

	Layer Attributes	SST-2	QNLI	MNLI	QQP	RTE
$\mathcal{P}_1$	Type	W_v	W_v, W_q	W_v, W_q	W_v, W_q	W_v, W_q
	Index	$\{21, \dots, 23\}$	$\{21, \dots, 23\}$	$\{21, \dots, 23\}$	$\{21, \dots, 23\}$	$\{21, \dots, 23\}$
$\mathcal{P}_2$	Type	W_v, W_q	W_v, W_q	W_v, W_q	W_v, W_q	W_v, W_q
	Index	$\{18, \dots, 23\}$	$\{15, \dots, 23\}$	$\{15, \dots, 23\}$	$\{15, \dots, 23\}$	$\{16, \dots, 23\}$
$\mathcal{P}_3$	Type	W_v, W_q	W_v, W_q	W_v, W_q	W_v, W_q	W_v, W_q
	Index	$\{0, \dots, 23\}$	$\{12, \dots, 23\}$	$\{12, \dots, 23\}$	$\{12, \dots, 23\}$	$\{12, \dots, 23\}$

Table 12: The selected layer set attached with LoRA for the setup of Figure 4

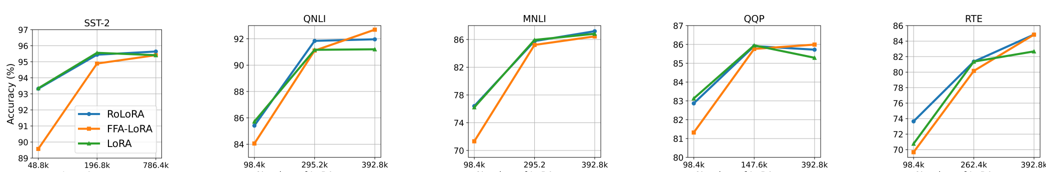

Figure 9: Results with RoBERTa-Large models on GLUE of different budget of finetuning parameters. It involves 3 clients using rank 8.

A5.5 Effect of Rank on RoLoRA and FFA-LoRA in the Centralized Setting

We evaluated FFA-LoRA and RoLoRA on MNLI and QQP using 8 LoRA adapters attached to query and value projection of last 4 layers, and trained for 50000 iterations to ensure full convergence. Increasing ranks can narrow the performance gap between the two schemes. Another related technique to narrow the performance gap between the two schemes is by increasing the number of adapters, as discussed in Section 5.1 ("Effect of Number of Fine-Tuning Parameters"). With sufficient adapters,

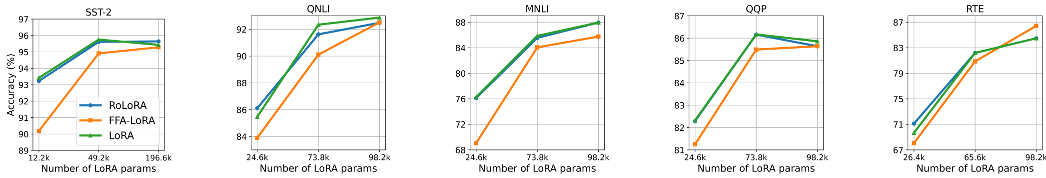

Figure 10: Results with RoBERTa-Large models on GLUE of different budget of finetuning parameters. It involves 3 clients using rank 2.

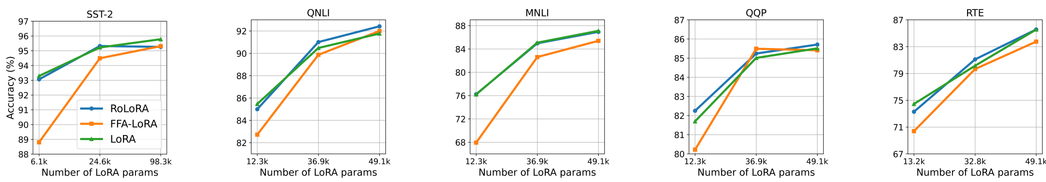

Figure 11: Results with RoBERTa-Large models on GLUE of different budget of finetuning parameters. It involves 3 clients using rank 1.

FFA-LoRA can achieve comparable peak accuracy to RoLoRA. However, in federated settings where resources are constraiend, RoLoRA is more advantageous.

		rank-1	rank-2	rank-32	rank-64
	FFA-LoRA	80.66	81.51	83.3	83.32
MNLI	RoLoRA	83.93	84.59	85.78	85.79
	Diff	3.27	+3.08	+2.51	+2.47
	FFA-LoRA	69.61	74.01	75.53	75.51
QQP	RoLoRA	77.26	77.41	78.03	78.05
	Diff	+7.65	+3.40	+2.5	+2.54

Table 13: Evaluation on FFA-LoRA and RoLoRA in the centralized setting using only 8 LoRA adapters. Increasing ranks can narrow the performance gap between the two schemes.

A5.6 Align the Communication Cost

In Figure 12, we conduct a comparison of three methods under the constraint of identical communication costs under the assumption that the number of clients is small. To align the communication costs across these methods, two approaches are considered. The first approach involves doubling the rank of FFA-LoRA and RoLoRA, with results presented in Appendix A5.4. The second approach requires doubling the number of layers equipped with LoRA modules. In Figure 12, the latter strategy is employed. Specifically, for both FFA-LoRA and RoLoRA, we adjust the communication costs by doubling the number of layers equipped with LoRA modules, thereby standardizing the size of the transmitted messages. The configurations are presented in Table 14. Figure 12 demonstrates that when operating within a constrained communication cost budget, the performance of RoLoRA surpasses that of the other two methods for most of the tasks.

In Table 14, we include the details about layers attached with LoRA adapters.

A5.7 Commonsense Reasoning Tasks

We present the configurations for Table 3 in Table 15. We show the results under the same setup but using rank-2 LoRA modules in Table 16.

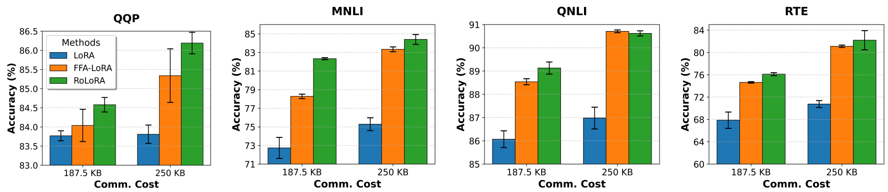

Figure 12: RoBERTa-Large accuracies on QQP, MNLI, QNLI, and RTE with specific uplink communication budget. It involves 3 clients using rank 4. Error bars reflect standard deviations.

	Communication Cost	LoRA	FFA-LoRA	RoLoRA
187.5 KB	Type	W_v, W_q	W_v, W_q	W_v, W_q
	Index	$\{21, \ldots, 23\}$	$\{18, \ldots, 23\}$	$\{18, \ldots, 23\}$
250 KB	Type	W_v, W_q	W_v, W_q	W_v, W_q
	Index	$\{20, \ldots, 23\}$	$\{16, \ldots, 23\}$	$\{16, \ldots, 23\}$

Table 14: The selected layer set attached with LoRA modules for the setup of Figure 12

A5.8 Language Generation Tasks

Model, Datasets and Metrics. We evaluate the performance of three federated finetuning methods with the model Llama-2-7B [35], on two datasets: CodeAlpaca [4] for coding tasks, and Alpaca [33] for general instruction-following tasks. Using HumanEval [5] as the metric for CodeAlpaca, we assess the model's ability to generate accurate code solutions. For Alpaca, we employ MMLU (Massive Multitask Language Understanding) [18] to evaluate the model's performance across diverse domains. This provides an assessment of Llama-2-7B's coding proficiency, and general language capabilities when finetuning in the federated setting. We show the setup in Table 17.

Results Table 18 presents the evaluation results of the Llama-2-7B model using three methods, across two tasks: HumanEval, and MMLU. The metrics reported include the average and standard deviation of performance over five seeds, with 50 clients involved. The results show that RoLoRA achieves the highest scores across most metrics, demonstrating slightly improved performance compared to LoRA and FFA-LoRA. The improvements are more evident in certain subcategories of the MMLU dataset.

A5.9 Communication and Time Cost Comparison

Table 19 compares the communication cost and time cost in a 50-client setting on MNLI task. RoLoRA and FFA-LoRA have the lowest communication and time costs.

A5.10 Privacy-preserving FL

Beyond robustness, practical FL often requires privacy-preserving mechanisms, most notably cryptographic approaches such as secure aggregation [3] or homomorphic encryption [19, 8, 7], and statistical protections like differential privacy (DP) [14, 41]. While RoLoRA was not explicitly designed with privacy mechanisms such as differential privacy, we recognize that its ability to mitigate inexactness through alternating optimization and aggregation may help mitigate a key challenge for DP-aware federated learning, where inexact model updates can be particularly problematic. In Table 20, we itegrated NbAFL [41] with $\epsilon = 10$ and $\delta = 1e-6$ in a 3-client setting. RoLoRA outperform others across two tasks.

Total comm. rounds	Batch size	Local Epochs	Layer type attached with LoRA	Layer index attached with LoRA
10	1	30	W_k, W_v, W_q, W_o	$\{26, \ldots, 31\}$

Table 15: Configurations for Commonsense Reasoning tasks.

	BoolQ	PIQA	SIQA	HellaSwag	WinoGrande	ARC-e	ARC-c	OBQA
LoRA	34.36$_{\pm 16.90}$	42.87$_{\pm 14.05}$	19.12$_{\pm 4.22}$	26.21$_{\pm 1.91}$	47.2$_{\pm 0.64}$	10.31$_{\pm 5.96}$	9.84$_{\pm 6.13}$	12.33$_{\pm 7.46}$
FFA-LoRA	44.04$_{\pm 11.48}$	51.46$_{\pm 9.81}$	25.38$_{\pm 11.27}$	23.86$_{\pm 2.67}$	46.93$_{\pm 1.54}$	22.25$_{\pm 7.92}$	20.65$_{\pm 6.33}$	20.67$_{\pm 5.33}$
RoLoRA	61.3$_{\pm 0.99}$	60.81$_{\pm 6.35}$	37.97$_{\pm 5.39}$	29.62$_{\pm 2.62}$	49.59$_{\pm 1.2}$	37.05$_{\pm 2.92}$	29.09$_{\pm 3.33}$	28.93$_{\pm 4.64}$

Table 16: Results with Llama-2-7B models on commonsense reasoning tasks. It involves 50 clients using rank 2.

Total comm. rounds	Batch size	Local Epochs	Layer type attached with LoRA	Layer index attached with LoRA
100	1	30	W_k, W_v, W_q, W_o	$\{24, \ldots, 31\}$

Table 17: Configurations for language generation tasks.

	LoRA	FFA-LoRA	RoLoRA
HumanEval	12.96$_{\pm 0.37}$	13.29$_{\pm 0.21}$	13.45$_{\pm 0.28}$
MMLU	45.81$_{\pm 0.03}$	45.80$_{\pm 0.02}$	45.93$_{\pm 0.01}$
human	43.26$_{\pm 0.04}$	43.24$_{\pm 0.01}$	43.46$_{\pm 0.02}$
other	52.67$_{\pm 0.06}$	52.72$_{\pm 0.05}$	52.83$_{\pm 0.04}$
social	51.73$_{\pm 0.04}$	51.64$_{\pm 0.05}$	51.81$_{\pm 0.04}$
stem	37.10$_{\pm 0.03}$	37.12$_{\pm 0.01}$	37.05$_{\pm 0.02}$

Table 18: Results with Llama-2-7B model on HumanEval, and MMLU. We report the average and std. over five seeds. The number of clients is 50. The metric used across all tasks is accuracy, where higher values indicate better performance.

Method	Comm. cost	Time cost
LoRA	1500.8 KB	0.0415 sec
FFA-LoRA	750.4 KB	0.0163 sec
FlexLoRA	1500.8 KB	2.252 sec
FLoRA	193603.2 KB	2.2179 sec
RoLoRA	750.4 KB	0.0182 sec

Table 19: Communication and Time Cost Comparison

Comm. cost: Total message size sent and received by each client per communication round in the MNLI experiment shown in Table 1.
Time cost: Mean server aggregation time per communication round for the MNLI experiment with 50 clients in Table 1, averaged over 30 runs.

	MNLI	QQP
LoRA	72.79 ± 5.23	57.97 ± 8.23
FFA-LoRA	79.65 ± 0.56	78.16 ± 0.6
RoLoRA	81.08 ± 0.81	81.64 ± 0.35

Table 20: Evaluation on differential privacy using $\epsilon = 10$ and $\delta = 1e - 6$ in a 3-client setting.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Please see Section 3, Section 4 and Section 5.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Appendix A2.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Please see Appendix A3 and Appendix A4.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See Appendix A5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The datasets are all open-source. The code is uploaded.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See Section 5 and Appendix A5

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: See standard deviation reported in Section 5 and Appendix A5

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please see Section 5

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We followed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Our paper discusses the broader societal implications of deploying federated learning methods enhanced with LoRA adapters. On the positive side, we emphasize that our approach can significantly improve the efficiency and scalability of federated learning, thereby facilitating model training on decentralized data. We did not identify any direct negative societal impacts, as our contribution is primarily methodological and does not involve deployment or data collection.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This paper proposes a framework for federated finetuning, and does not pose high risks for misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer:[Yes]

Justification: This paper is licensed under CC-BY-NC-SA 4.0. All other codes, datasets, and references are properly cited.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: We did not have any released new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We did not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: We did not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.