LoRA-FAIR: Federated LoRA Fine-Tuning with Aggregation and Initialization Refinement

Supplementary Material

9. Additional Experiments Results

In this section, we provide additional experimental details and results to further validate our proposed method, LoRA-FAIR.

9.1. Convergence Performance

We present the convergence performance of our proposed method compared to baseline methods using the ViT or MLP-Mixer model under feature non-IID setting. As shown in Fig. 7 and Fig. 8, our proposed method consistently outperforms all baseline methods. These results are consistent with those in the main paper (Tab. 2), further validating the robustness of our approach.

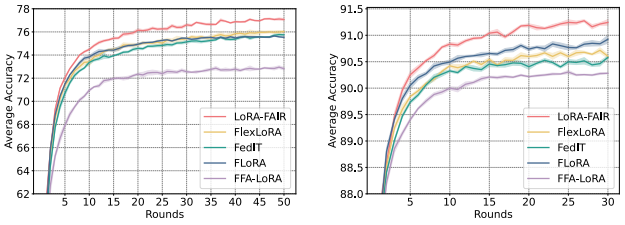

Figure 7. **Comparison of average accuracy** across training rounds on DomainNet (**left**) and NICO++ (**right**) datasets using the ViT model. The shaded area indicates the variance across multiple runs. For more details, refer to Sec. 5.1.

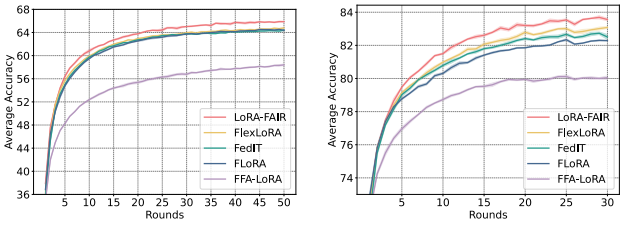

Figure 8. **Comparison of average accuracy** across training rounds on DomainNet (**left**) and NICO++ (**right**) datasets using the Mixer model. The shaded area indicates the variance across multiple runs. For more details, refer to Sec. 5.1.

9.2. Adaptation for Clients with Heterogeneous LoRA Ranks

Our proposed method primarily focuses on settings where clients have the same LoRA rank, addressing challenges such as server aggregation bias and client initialization lag when combining LoRA with federated learning. However, our approach can be extended to scenarios where clients have heterogeneous LoRA ranks.

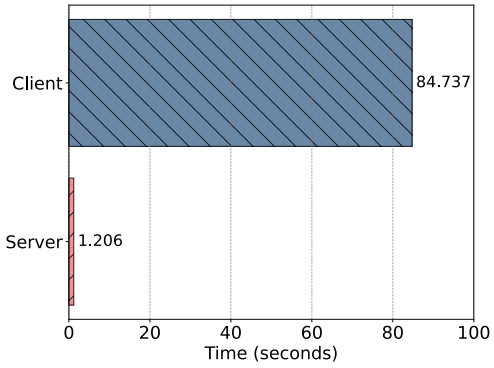

Figure 9. **Comparison of computational time** between the client and the server. See details in Sec. 9.3

The state-of-the-art method for handling heterogeneous ranks in FL is HETLoRA [7], which employs zero-padding and truncation for distribution. It is important to note that HETLoRA is specifically designed for heterogeneous settings and operates orthogonally to our proposed method. By integrating zero-padding and truncation for distribution into LoRA-FAIR, our method can effectively operate in heterogeneous rank settings. We evaluate this adaptation using the DomainNet dataset with ViT as the foundation model. The client data distribution and training settings are consistent with those used in the feature non-IID experiments in the main paper. The clients LoRA ranks are set as $\{2, 4, 4, 6, 6, 8\}$. The results, presented in Tab. 6, demonstrate that our proposed method, combined with zero-padding and truncation, achieves the best performance compared to existing methods, validating its effectiveness in heterogeneous rank scenarios. We note that due to the heterogeneous LoRA ranks, FedIT and FFA-LoRA are not suitable for this setting and are therefore excluded from the experiment. While FLoRA can operate under heterogeneous settings, it is not included in the results as it fails to converge in our experiments. This failure underscores its limitation of directly adding updates to the pre-trained model rather than updating the LoRA modules.

9.3. Server-Side Computational Overhead Analysis

Our proposed method addresses both server aggregation bias and client initialization lag by solving Eq. (8), introducing only a small computational overhead on the server side. In our main experiment, we solve Eq. (8) using SGD with a learning rate of 0.01 and 1000 iterations. However,

Method	Clipart	Infograph	Painting	Quickdraw	Real	Sketch	Average
HETLoRA	73.96	42.57	74.49	27.52	87.05	59.74	60.89
FlexLoRA	83.11	52.43	78.63	62.30	88.23	77.32	73.50
LoRA-FAIR + HETLoRA	**83.40**	**52.25**	**79.28**	**63.24**	**89.40**	**77.74**	**74.22**

Table 6. **Performance comparison** with baselines across different domains on DomainNet using ViT model with client having heterogeneous LoRA rank. **Average** means the average accuracy across all domains. See details in Sec. 9.2.

Method	Local Epoch	Clipart	Infograph	Painting	Quickdraw	Real	Sketch	Average	Δ
FLoRA	2	85.15	53.51	79.43	70.09	89.25	77.20	75.53	-
Proposed	2	86.25	56.26	80.09	71.25	89.52	79.06	77.07	+1.54
FLoRA	10	83.81	52.91	78.36	61.25	88.68	76.57	73.60	-
Proposed	10	85.20	53.39	79.03	61.51	89.24	77.47	74.31	+0.71

Table 7. **Limitation of FLoRA's Reinitialization**. FLoRA's reinitialization strategy fails to learn an optimal client update under smaller local training, leading to suboptimal model performance. See Sec. 9.4 for details.

this additional cost is minimal and can be considered negligible given the substantial computational resources typically available on servers. Moreover, a comparison of training times, as shown in Fig. 9, demonstrates that the time required to solve Eq. (8) on the server is minimal compared to the client-side local training time.

9.4. Limitation of FLoRA's Reinitialization

We evaluate performance under different local epochs. In our main experiments with the feature non-IID setting, we set the number of local epochs to 2 and the number of global rounds to 50. Here, we test a configuration where the local epochs are set to 10, and accordingly, the global rounds are reduced to 10 to maintain a fixed total number of updates. The results indicate that a shorter local epoch with more frequent updates leads to better performance for both the proposed method and FLoRA. This finding is consistent with [48], which suggests that the number of local epochs should not be too high in a non-IID FL setting. Furthermore, with shorter local epochs, the performance gap between our proposed method and FLoRA increases, further validating that FLoRA's reinitialization strategy fails to learn an optimal client update under limited local training, ultimately degrading the final model performance.

10. Prior Works

In this section, we review existing methods and their limitations.

FedIT [46]: FedIT is the earliest approach to integrate LoRA with FedAvg. In FedIT, each client starts with a fixed pre-trained foundation model and trains local LoRA modules, represented as low-rank matrices $\mathbf{A}_k$ and $\mathbf{B}_k$, on its private dataset. The server aggregates these local matrices into global LoRA modules through a weighted average based on data size. While computationally efficient, this method introduces server-side aggregation bias.

FFA-LoRA [33]: FFA-LoRA freezes the non-zero initialized low-rank matrix $\mathbf{A}$ and updates only the zero-initialized matrix $\mathbf{B}$. By freezing $\mathbf{A}$, the actual global update becomes equal to the ideal global update (i.e., $\mathbf{\Delta W} = \mathbf{\Delta W}'$), addressing server-side aggregation bias. However, freezing $\mathbf{A}$ significantly reduces the number of trainable parameters, limiting the model's capacity. Our experiments confirm that although FFA-LoRA resolves aggregation bias, its limited parameter flexibility results in worse performance compared to other baselines.

FLoRA [42]: FLoRA stacks local LoRA modules from all clients and transmits the stacked modules back to each client to reconstruct global updates, which are then added directly to each client's pre-trained model while reinitializing local LoRA modules for the next training round. Although FLoRA effectively addresses server-side aggregation bias, it incurs high communication costs proportional to the number of clients and raises privacy concerns, as it distributes all clients' LoRA modules rather than only the averaged ones. Additionally, FLoRA's reinitialization strategy introduces Client-Side Initialization Lag. Frequent reinitialization results in small gradient updates, leading to inefficient training and suboptimal performance.

FlexLoRA [2]: FlexLoRA reformulates each client's local LoRA modules into a local update, sums these updates to generate a global update, and applies SVD to produce global LoRA modules. These modules are then distributed to clients as initialization for the next round. While this approach formulates an ideal global update, it still suffers from server-side aggregation bias due to the SVD step. For example, consider two clients, each with rank-8 LoRA modules ($\text{Rank}(\Delta W_1) = 8$ and $\text{Rank}(\Delta W_2) = 8$), resulting in a global update with $\text{Rank}(\Delta W) \leq 16$. Using SVD to produce global modules with a rank of 8 may lead to information loss, preventing the transmission of an ideal global update to clients.

Comparison to Existing Efficient Weight Aggregation FL Methods. Our work identifies a gap in existing federated learning methods concerning fine-tuning with LoRA. While prior approaches—such as layer-wise model aggregation [25], elastic aggregation [6], and related layer-wise techniques [18, 30]—have demonstrated effectiveness in general federated optimization, they are not well-suited for

LoRA-based fine-tuning. Specifically, these methods do not address how to decompose aggregated model updates into the necessary LoRA modules for client model initialization, nor do they provide strategies to avoid the direct transmission of these updates. To overcome these limitations, our method tailors the aggregation process specifically for federated fine-tuning with LoRA, bridging the gap left by existing techniques.

11. Theorem

Theorem 11.1. *For analytical tractability, we consider the case where the similarity metric S is based on the Frobenius norm. The residual correction term $\Delta\mathbf{B}$ obtained by minimizing Equation (8) guarantees that $(\bar{\mathbf{B}} + \Delta\mathbf{B})\bar{\mathbf{A}}$ approaches the ideal global update $\Delta\mathbf{W}$ with the following approximation guarantee:*

$$\begin{aligned}&\|(\bar{\mathbf{B}} + \Delta\mathbf{B}^*)\bar{\mathbf{A}} - \Delta\mathbf{W}\|_F^2 \\ \leq&\|\Delta\mathbf{W} - \bar{\mathbf{B}}\bar{\mathbf{A}}\|_F^2 \cdot \left(1 - \frac{\sigma_{\min}^2(\bar{\mathbf{A}})}{\sigma_{\min}^2(\bar{\mathbf{A}}) + \lambda}\right)^2, \end{aligned} \tag{9}$$

here $\sigma_{\min}(\bar{\mathbf{A}})$ is the smallest non-zero singular value of $\bar{\mathbf{A}}$.

Proof. Let's denote $\mathbf{E} = \Delta\mathbf{W} - \bar{\mathbf{B}}\bar{\mathbf{A}}$ as the initial aggregation error. Our objective function becomes:

$$J(\Delta\mathbf{B}) = \|\Delta\mathbf{B}\bar{\mathbf{A}} - \mathbf{E}\|_F^2 + \lambda\|\Delta\mathbf{B}\|_F^2 \tag{10}$$

To find the critical points of $J(\Delta\mathbf{B})$, we take the derivative with respect to $\Delta\mathbf{B}$ and set it equal to zero:

$$\nabla_{\Delta\mathbf{B}} J(\Delta\mathbf{B}) = 2(\Delta\mathbf{B}\bar{\mathbf{A}} - \mathbf{E})\bar{\mathbf{A}}^T + 2\lambda\Delta\mathbf{B} = 0 \tag{11}$$

Solving for the optimal $\Delta\mathbf{B}^*$:

$$\Delta\mathbf{B}^* = \mathbf{E}\bar{\mathbf{A}}^T(\bar{\mathbf{A}}\bar{\mathbf{A}}^T + \lambda\mathbf{I})^{-1}, \tag{12}$$

where $\mathbf{I}$ is the identity matrix of appropriate dimensions. Substituting back the definition of $\mathbf{E}$:

$$\Delta\mathbf{B}^* = (\Delta\mathbf{W} - \bar{\mathbf{B}}\bar{\mathbf{A}})\bar{\mathbf{A}}^T(\bar{\mathbf{A}}\bar{\mathbf{A}}^T + \lambda\mathbf{I})^{-1} \tag{13}$$

The residual error after applying the correction is:

$$\begin{aligned}\mathbf{E}_{residual} &= -\mathbf{E} + \Delta\mathbf{B}^*\bar{\mathbf{A}} \\ &= -\mathbf{E} + \mathbf{E}\bar{\mathbf{A}}^T(\bar{\mathbf{A}}\bar{\mathbf{A}}^T + \lambda\mathbf{I})^{-1}\bar{\mathbf{A}} \\ &= \mathbf{E}(-\mathbf{I} + \bar{\mathbf{A}}^T(\bar{\mathbf{A}}\bar{\mathbf{A}}^T + \lambda\mathbf{I})^{-1}\bar{\mathbf{A}}) \end{aligned} \tag{14}$$

Let's define the matrix $\mathbf{M} = -\mathbf{I} + \bar{\mathbf{A}}^T(\bar{\mathbf{A}}\bar{\mathbf{A}}^T + \lambda\mathbf{I})^{-1}\bar{\mathbf{A}}$. For any matrix $\bar{\mathbf{A}}$, the eigenvalues of $\bar{\mathbf{A}}^T(\bar{\mathbf{A}}\bar{\mathbf{A}}^T + \lambda\mathbf{I})^{-1}\bar{\mathbf{A}}$ can be bounded using the properties of matrix norms and the Sherman-Morrison-Woodbury formula:

$$\bar{\mathbf{A}}^T(\bar{\mathbf{A}}\bar{\mathbf{A}}^T + \lambda\mathbf{I})^{-1}\bar{\mathbf{A}} = \bar{\mathbf{A}}^T\bar{\mathbf{A}}(\bar{\mathbf{A}}^T\bar{\mathbf{A}} + \lambda\mathbf{I})^{-1} \tag{15}$$

The eigenvalues of this matrix are of the form $\frac{\mu_i}{\mu_i+\lambda}$, where μ_i are the eigenvalues of $\bar{\mathbf{A}}^T\bar{\mathbf{A}}$. Since the eigenvalues of $\bar{\mathbf{A}}^T\bar{\mathbf{A}}$ are the squares of the singular values of $\bar{\mathbf{A}}$, i.e., $\mu_i = \sigma_i^2$, the eigenvalues of $\bar{\mathbf{A}}^T(\bar{\mathbf{A}}\bar{\mathbf{A}}^T + \lambda\mathbf{I})^{-1}\bar{\mathbf{A}}$ are $\frac{\sigma_i^2}{\sigma_i^2+\lambda}$. Therefore, the eigenvalues of $\mathbf{M} = -\mathbf{I} + \bar{\mathbf{A}}^T(\bar{\mathbf{A}}\bar{\mathbf{A}}^T + \lambda\mathbf{I})^{-1}\bar{\mathbf{A}}$ are $-1 + \frac{\sigma_i^2}{\sigma_i^2+\lambda} = -\frac{\lambda}{\sigma_i^2+\lambda}$.

The spectral norm of $\mathbf{M}$ is the maximum absolute eigenvalue:

$$\|\mathbf{M}\|_2 = \max_i \left|\frac{-\lambda}{\sigma_i^2 + \lambda}\right| = \frac{\lambda}{\sigma_{\min}^2 + \lambda} \tag{16}$$

Using the property that for any matrices $\mathbf{P}$ and $\mathbf{Q}$, $\|\mathbf{PQ}\|_F \leq \|\mathbf{P}\|_F\|\mathbf{Q}\|_2$, we have:

$$\|\mathbf{E}_{residual}\|_F = \|\mathbf{EM}\|_F \tag{17}$$

$$\leq \|\mathbf{E}\|_F\|\mathbf{M}\|_2 \tag{18}$$

$$= \|\mathbf{E}\|_F \cdot \frac{\lambda}{\sigma_{\min}^2 + \lambda} \tag{19}$$

Since $\frac{\lambda}{\sigma_{\min}^2+\lambda} = 1 - \frac{\sigma_{\min}^2}{\sigma_{\min}^2+\lambda}$, we have:

$$\|\mathbf{E}_{residual}\|_F \leq \|\mathbf{E}\|_F \cdot \left(1 - \frac{\sigma_{\min}^2}{\sigma_{\min}^2 + \lambda}\right) \tag{20}$$

Squaring both sides and substituting $\mathbf{E} = \Delta\mathbf{W} - \bar{\mathbf{B}}\bar{\mathbf{A}}$, we get our final bound.

□

Corollary 11.2. *When $\lambda = 0$ and $\bar{\mathbf{A}}$ has full row rank, there exists an exact solution where:*

$$(\bar{\mathbf{B}} + \Delta\mathbf{B}^*)\bar{\mathbf{A}} = \Delta\mathbf{W}$$

As the regularization parameter λ increases, the solution balances two objectives:

1. Minimizing the approximation error to the ideal update $\Delta\mathbf{W}$
2. Preventing large deviations from the averaged LoRA module $\bar{\mathbf{B}}$

Theorem 11.3 (Convergence of Federated LoRA Fine–Tuning)**.** *Under standard FL assumptions (L-smooth loss, bounded gradients G, E local epochs), and assuming the*

global learning rate η, the convergence of federated LoRA fine-tuning after T rounds is:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\mathcal{L}(W^t)\|^2] \leq \frac{4[\mathcal{L}(W^0)-\mathcal{L}(W^*)]}{\eta T} + 4\eta^2E^2G^2\left(\frac{L^2}{2}+1\right) + 8\frac{1}{T}\sum_{t=0}^{T-1}\|\Delta W^t - \bar{B}^t\bar{A}^t\|_F^2\cdot\gamma$$

where γ characterizes the aggregation method: 1. For LoRA-FAIR: $\gamma = \left(1-\frac{\sigma^2_{\min}(\bar{A}^t)}{\sigma^2_{\min}(\bar{A}^t)+\lambda}\right)^2 < 1$. 2. For FedIT (standard aggregation): $\gamma = 1$

Proof. At round t, the global model update is:

$$W^{t+1} = W^t - \eta\cdot\Delta W'^t \tag{21}$$

where $\Delta W'^t = (\bar{B}^t + \Delta B^{*t})\bar{A}^t$ for LoRA-FAIR and $\Delta W'^t = \bar{B}^t\bar{A}^t$ for FedIT. The ideal global update is:

$$\Delta W^t = \sum_{k=1}^{K} p_k B_k^t A_k^t \tag{22}$$

Define the aggregation error:

$$E_{agg}^t = \Delta W^t - \Delta W'^t \tag{23}$$

From Theorem A.1, for LoRA-FAIR:

$$\|E_{agg}^t\|_F^2 \leq \|\Delta W^t - \bar{B}^t\bar{A}^t\|_F^2\cdot\gamma \tag{24}$$

Using L-smoothness:

$$\begin{aligned}\mathcal{L}(W^{t+1}) &\leq \mathcal{L}(W^t) + \langle\nabla\mathcal{L}(W^t), W^{t+1}-W^t\rangle \\ &\quad + \frac{L}{2}\|W^{t+1}-W^t\|^2 \qquad (25)\\ &= \mathcal{L}(W^t) - \eta\langle\nabla\mathcal{L}(W^t), \Delta W'^t\rangle \\ &\quad + \frac{L\eta^2}{2}\|\Delta W'^t\|^2 \qquad (26)\end{aligned}$$

Decomposing $\Delta W'^t = \Delta W^t - E_{agg}^t$:

$$\begin{aligned}\langle\nabla\mathcal{L}(W^t), \Delta W'^t\rangle &= \langle\nabla\mathcal{L}(W^t), \Delta W^t\rangle \\ &\quad - \langle\nabla\mathcal{L}(W^t), E_{agg}^t\rangle \qquad (27)\end{aligned}$$

Under bounded gradients:

$$\langle\nabla\mathcal{L}(W^t), \Delta W^t\rangle \geq \frac{1}{2}\|\nabla\mathcal{L}(W^t)\|^2 - \frac{\eta^2E^2L^2G^2}{2} \tag{28}$$

For the error term:

$$\langle\nabla\mathcal{L}(W^t), -E_{agg}^t\rangle \leq \frac{1}{4}\|\nabla\mathcal{L}(W^t)\|^2 + \|E_{agg}^t\|^2 \tag{29}$$

Combining bounds:

$$\begin{aligned}\mathcal{L}(W^{t+1}) &\leq \mathcal{L}(W^t) - \frac{\eta}{4}\|\nabla\mathcal{L}(W^t)\|^2 + \frac{\eta^3E^2L^2G^2}{2} \\ &\quad + \eta\|E_{agg}^t\|^2 + \frac{L\eta^2}{2}\|\Delta W'^t\|^2 \qquad (30)\end{aligned}$$

Using $\|\Delta W'^t\|^2 \leq 2(\|\Delta W^t\|^2 + \|E_{agg}^t\|^2)$ and the bound on E_{agg}^t:

$$\begin{aligned}\mathcal{L}(W^{t+1}) &\leq \mathcal{L}(W^t) - \frac{\eta}{4}\|\nabla\mathcal{L}(W^t)\|^2 + \frac{\eta^3E^2L^2G^2}{2} \\ &\quad + \eta\|E_{agg}^t\|^2 + \frac{L\eta^2}{2}(2(\|\Delta W^t\|^2 + \|E_{agg}^t\|^2)) \\ &\leq \mathcal{L}(W^t) - \frac{\eta}{4}\|\nabla\mathcal{L}(W^t)\|^2 + \frac{\eta^3E^2L^2G^2}{2} \\ &\quad + \eta\|E_{agg}^t\|^2 + \eta(\eta^2E^2G^2 + \|E_{agg}^t\|^2) \\ &\leq \mathcal{L}(W^t) - \frac{\eta}{4}\|\nabla\mathcal{L}(W^t)\|^2 + \frac{\eta^3E^2L^2G^2}{2} \\ &\quad + 2\eta\|E_{agg}^t\|^2 + \eta^3E^2G^2 \qquad (31)\end{aligned}$$

where $\eta < \frac{1}{L}$. Then we have:

$$\begin{aligned}\mathcal{L}(W^{t+1}) &\leq \mathcal{L}(W^t) - \frac{\eta}{4}\|\nabla\mathcal{L}(W^t)\|^2 + \frac{\eta^3E^2L^2G^2}{2} \\ &\quad + 2\eta\|\Delta W^t - \bar{B}^t\bar{A}^t\|_F^2\cdot\gamma + \eta^3E^2G^2 \qquad (32)\end{aligned}$$

Rearranging and taking expectation:

$$\begin{aligned}\frac{\eta}{4}\mathbb{E}[\|\nabla\mathcal{L}(W^t)\|^2] &\leq \mathbb{E}[\mathcal{L}(W^t)] - \mathbb{E}[\mathcal{L}(W^{t+1})] \\ &\quad + \frac{\eta^3E^2L^2G^2}{2} + \eta^3E^2G^2 \\ &\quad + 2\eta\|\Delta W^t - \bar{B}^t\bar{A}^t\|_F^2\cdot\gamma \qquad (33)\end{aligned}$$

Then by summing over t:

$$\begin{aligned}\frac{\eta}{4}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\mathcal{L}(W^t)\|^2] &\leq \mathcal{L}(W^0) - \mathbb{E}[\mathcal{L}(W^T)] \\ &\quad + T\eta^3E^2G^2(\frac{L^2}{2}+1) \\ &\quad + 2\eta\sum_{t=0}^{T-1}\|\Delta W^t - \bar{B}^t\bar{A}^t\|_F^2\cdot\gamma \qquad (34)\end{aligned}$$

dividing both sides by $\frac{\eta}{4}$ and T, we have:

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\mathcal{L}(W^t)\|^2] &\leq \frac{4}{T\eta}[\mathcal{L}(W^0) - \mathcal{L}(W^*)] \\ &\quad + 4\eta^2E^2G^2(\frac{L^2}{2}+1) \\ &\quad + 8\frac{1}{T}\sum_{t=0}^{T-1}\|\Delta W^t - \bar{B}^t\bar{A}^t\|_F^2\cdot\gamma \qquad (35)\end{aligned}$$

With $\eta = \frac{1}{\sqrt{T}}$, we finally get:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla\mathcal{L}(W^t)\|^2] \leq \frac{4[\mathcal{L}(W^0)-\mathcal{L}(W^*)]}{\sqrt{T}} + 4\eta^2E^2G^2(\frac{L^2}{2}+1) + 8\frac{1}{T}\sum_{t=0}^{T-1}\|\Delta W^t - \bar{B}^t\bar{A}^t\|_F^2\cdot\gamma \tag{36}$$

□

As data becomes more non-IID, $\|\Delta\mathbf{W} - \bar{\mathbf{B}}\bar{\mathbf{A}}\|_F$ increases, leading to a larger error term in standard methods. LoRA-FAIR reduces this impact by the factor $\gamma < 1$, providing tighter convergence bounds especially in highly non-IID settings.