Supplementary Material for the Paper: “Wireless Personalized Federated Fine-Tuning of Large Language Models via Low-Rank Adaptation”

Haofeng Sun, Hui Tian, Jingheng Zheng, and Wanli Ni

Appendix A
Proof of Theorem 1

With Assumptions 1, 2, 3, we investigate the convergence behavior of global loss function (2) composed by the desired local loss function $\nabla F_k(\mathbf{W}_{t,n})$ in the following theorem.

Theorem 2: *Suppose that Assumption 1 holds. The convergence behavior of the global loss function with respect to multiple tasks in AirPFed-LoRA framework between two consecutive fine-tuning rounds can be expressed by*

$$
\begin{aligned}
&\sum\nolimits_{n\in\mathcal{N}}\frac{D_n}{D}\left(F(\{\mathbf{W}_{t+1,n}\})-F(\{\mathbf{W}_{t,n}\})\right)\\
&\leq \underbrace{\mathrm{Vec}(\mathbf{W}^s_{t+1}-\mathbf{W}^s_t)^{\mathrm{T}}\mathrm{Vec}(\nabla_{\mathbf{W}^s_t}F\left(\{\mathbf{W}_{t,n}\}\right))}_{E_{t,1}}\\
&+\underbrace{\sum_{n\in\mathcal{N}}\frac{D_n}{D}\mathrm{Vec}(\mathbf{W}^p_{t+1,n}-\mathbf{W}^p_{t,n})^{\mathrm{T}}\mathrm{Vec}(\nabla_{\mathbf{W}^p_{t,n}}F^E_n(\mathbf{W}_{t,n}))}_{E_{t,2}}\\
&+\frac{\beta}{2}\underbrace{\|\mathbf{W}^s_{t+1}-\mathbf{W}^s_t\|^2_{\mathrm{F}}}_{E_{t,3}}+\underbrace{\frac{\beta}{2}\sum_{n\in\mathcal{N}}\frac{D_n}{D}\|\mathbf{W}^p_{t+1,n}-\mathbf{W}^p_{t,n}\|^2_{\mathrm{F}}}_{E_{t,4}},
\end{aligned}
\tag{38}
$$

where $\mathbf{W}^s_t=[\mathbf{B}^s_t;\mathbf{A}^s_t]$ *denotes the s-adapter matrix,* $\mathbf{W}^p_{t,n}=[\mathbf{B}^p_{t,n};\mathbf{A}^p_{t,n}]$ *denotes the p-adapter matrix of task* n*, and* $F^E_n(\mathbf{W}_{t,n})=\sum_{k\in\mathcal{K}_n}\frac{D_k}{D_n}F_k(\mathbf{W}_{t,n})$ *denotes the desired edge loss function of task* n*.*

Theorem 2 demonstrates that the convergence behavior of the global loss function is affected by four terms, i.e., $E_{t,1}$, $E_{t,2}$, $E_{t,3}$ and $E_{t,4}$. With Assumptions 2 and 3, we further bound the expectations of $E_{t,1}+E_{t,2}$ and $E_{t,3}+E_{t,4}$ in Lemma 1 and 2, respectively.

Lemma 1: *The upper bound of* $\mathbb{E}\left[E_{t,1}+E_{t,2}\right]$ *is given by*

$$
\begin{aligned}
\mathbb{E}\left[E_{t,1}+E_{t,2}\right]&\leq-\frac{\eta}{2}\left\|\sum_{n\in\mathcal{N}}\sum_{k\in\mathcal{K}_n}\frac{D_k}{D}\nabla F_k(\mathbf{W}_{t,n})\right\|^2_{\mathrm{F}}\\
&-\frac{\eta}{2}\left\|\nabla_{\mathbf{W}^s_t}F\left(\{\mathbf{W}_{t,n}\}\right)\right\|^2_{\mathrm{F}}-\frac{\eta}{2}\sum_{n\in\mathcal{N}}\frac{D_n}{D}\left\|\nabla_{\mathbf{W}^p_{t,n}}F^E_n(\mathbf{W}_{t,n})\right\|^2_{\mathrm{F}}.
\end{aligned}
\tag{39}
$$

Lemma 2: *Suppose that Assumptions 2 and 3 hold. The expectation of* $E_{t,3}+E_{t,4}$ *can be bounded by*

$$
\begin{aligned}
\mathbb{E}[E_{t,3}+E_{t,4}]&\leq\lambda\eta^2(C+1)\left\|\sum_{n\in\mathcal{N}}\sum_{k\in\mathcal{K}_n}\frac{D_k}{D}\nabla F_k(\mathbf{W}_{t,n})\right\|^2_{\mathrm{F}}\\
&+\eta^2(C+1)\sigma^2_{\text{dif}}+\eta^2\sigma^2_{\text{SGD}}+\eta^2\left\|\hat{\mathbf{G}}^{\mathbf{W}^s_t}_t-\mathbf{G}^{\mathbf{W}^s_t}_t\right\|^2_{\mathrm{F}}\\
&+\eta^2\sum\nolimits_{n\in\mathcal{N}}\frac{D_n}{D}\left\|\hat{\mathbf{G}}^{\mathbf{W}^p_{t,n}}_{t,n}-\mathbf{G}^{\mathbf{W}^p_{t,n}}_{t,n}\right\|^2_{\mathrm{F}}
\end{aligned}
\tag{40}
$$

By leveraging Theorem 2, Lemma 1 and 2, we derive an optimality gap of the AirPFed-LoRA framework and establish an upper bound for the variations of s- and p-adapters' matrices in Theorem 3.

Theorem 3: *Building upon Theorem 2, Lemma 1 and 2, while setting* $\eta\leq\frac{1}{\lambda(C+1)\beta}$*, the optimality gap of the AirPFed-LoRA framework between two consecutive fine-tuning rounds can be bounded by*

$$
\begin{aligned}
&\sum\nolimits_{n\in\mathcal{N}}\frac{D_n}{D}\left(F(\{\mathbf{W}_{t+1,n}\})-F(\{\mathbf{W}_{t,n}\})\right)\leq-\frac{\eta}{2}\Lambda_t\\
&+\frac{\beta\eta^2}{2}\sigma^2_{\text{SGD}}+\frac{\beta\eta^2}{2}(C+1)\sigma^2_{\text{dif}}+\frac{\beta\eta^2(L+Q)R}{2}\sigma^2_t\Pi_t
\end{aligned}
\tag{41}
$$

where $\Pi_t=\frac{R^s}{R}\mathrm{MSE}^s_t+\frac{R^p}{R}\sum_{n\in\mathcal{N}}\frac{D_n}{D}\mathrm{MSE}^p_{t,n}$ *denotes the combined MSE of the edge and global aggregated gradient signals in AirComp,* $R=R^s+R^p$ *denotes the total rank of s- and p-adapters,* Λ_t *is the F-norm of gradient matrices which captures the variation of s- and p-adapters, given by*

$$
\Lambda_t=\left\|\nabla_{\mathbf{W}^s_t}F\left(\{\mathbf{W}_{t,n}\}\right)\right\|^2_{\mathrm{F}}+\sum_{n\in\mathcal{N}}\frac{D_n}{D}\left\|\nabla_{\mathbf{W}^p_{t,n}}F^E_n(\mathbf{W}_{t,n})\right\|^2_{\mathrm{F}}.
\tag{42}
$$

Suppose that Assumptions 1, 2, 3 hold while set the learning rate as $\eta=\frac{1}{\lambda(C+1)\beta}$*. Then, the average gradients variation* Λ_t *can be bounded by*

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\Lambda_t&\leq\frac{2\sum_{n\in\mathcal{N}}\frac{D_n}{D}\left(F\left(\{\mathbf{W}_{1,n}\}\right)-F\left(\{\mathbf{W}^*_n\}\right)\right)}{\eta T}\\
&+\frac{\sigma^2_{\text{SGD}}}{\lambda(C+1)}+\frac{\sigma^2_{\text{dif}}}{\lambda}+\frac{(L+Q)R}{\lambda(C+1)T}\sum\nolimits_{t=1}^{T}\sigma^2_t\Pi_t=\Psi_T,
\end{aligned}
\tag{43}
$$

where $\{\mathbf{W}^*_n\}$ *denotes the optimal global s-adapter and edge p-adapters matrices.*

H. Sun, H. Tian, and J. Zheng are with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: sunhaofeng@bupt.edu.cn; tianhui@bupt.edu.cn; zhengjh@bupt.edu.cn).

W. Ni is with the Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: niwanli@tsinghua.edu.cn).