Convergence Analysis

We first clarify some notations used for convergence derivation. $t \in \{0,\ldots,T-1\}$ is the t-th communication round. $e \in \{0,1,\ldots,E\}$ is the e-th local iteration. $tE+0$ denotes the start of the $(t+1)$-th round in which client k in the $(t+1)$-th round receives the small homogeneous feature extractor $\mathcal{G}(\theta^t)$ from the server. $tE+e$ is the e-th local iteration in the $(t+1)$-th round. $tE+E$ is the last local iteration in the $(t+1)$-th round. After that, client k sends its local updated small homogeneous feature extractor the server for aggregation. $\mathcal{H}_k(h_k)$ is client k's entire local model consisting of the global small homogeneous feature extractor $\mathcal{G}(\theta)$ and the local heterogeneous model $\mathcal{F}_k(\omega_k)$ weighed by the trainable weight vector $\boldsymbol{\alpha}_k$, *i.e.*, $\mathcal{H}_k(h_k) = (\mathcal{G}(\theta) \circ \mathcal{F}_k(\omega_k)|\boldsymbol{\alpha}_k)$. η is the learning rate of client k's local model $\mathcal{H}_k(h_k)$, consisting of $\{\eta_\theta, \eta_\omega, \eta_{\boldsymbol{\alpha}}\}$.

Assumption 1: **Lipschitz Smoothness**. The gradients of client k's entire local heterogeneous model h_k are $L1$–Lipschitz smooth [47],

$$\begin{gathered}\|\nabla\mathcal{L}_k^{t_1}(h_k^{t_1};\boldsymbol{x},y)-\nabla\mathcal{L}_k^{t_2}(h_k^{t_2};\boldsymbol{x},y)\| \leqslant L_1\|h_k^{t_1}-h_k^{t_2}\|,\\ \forall t_1,t_2>0, k\in\{0,1,\ldots,N-1\},(\boldsymbol{x},y)\in D_k.\end{gathered} \tag{13}$$

The above formulation can be re-expressed as:

$$\mathcal{L}_k^{t_1}-\mathcal{L}_k^{t_2} \leqslant \langle\nabla\mathcal{L}_k^{t_2},(h_k^{t_1}-h_k^{t_2})\rangle+\frac{L_1}{2}\|h_k^{t_1}-h_k^{t_2}\|_2^2. \tag{14}$$

Assumption 2: **Unbiased Gradient and Bounded Variance**. Client k's random gradient $g_{h,k}^t=\nabla\mathcal{L}_k^t(h_k^t;\mathcal{B}_k^t)$ ($\mathcal{B}$ is a batch of local data) is unbiased,

$$\mathbb{E}_{\mathcal{B}_k^t\subseteq D_k}[g_{h,k}^t]=\nabla\mathcal{L}_k^t(h_k^t), \tag{15}$$

and the variance of random gradient $g_{h,k}^t$ is bounded by:

$$\mathbb{E}_{\mathcal{B}_k^t\subseteq D_k}[\|\nabla\mathcal{L}_k^t(h_k^t;\mathcal{B}_k^t)-\nabla\mathcal{L}_k^t(h_k^t)\|_2^2]\leqslant\sigma^2. \tag{16}$$

Assumption 3: **Bounded Parameter Variation**. The parameter variations of the small homogeneous feature extractor θ_k^t and θ^t before and after aggregation are bounded by:

$$\|\theta^t-\theta_k^t\|_2^2\leq\delta^2. \tag{17}$$

Based on the above assumptions, we can derive the following Lemma and Theorem.

Lemma 1: **Local Training.** Given Assumptions 1 and 2, the loss of an arbitrary client's local model h in the $(t+1)$-th local training round is bounded by:

$$\begin{aligned}\mathbb{E}[\mathcal{L}_{(t+1)E}] \leq \mathcal{L}_{tE+0}+(\frac{L_1\eta^2}{2}-\eta)\sum_{e=0}^{E}\|\nabla\mathcal{L}_{tE+e}\|_2^2\\ +\frac{L_1E\eta^2\sigma^2}{2}.\end{aligned} \tag{18}$$

Proof 1: An arbitrary client k's local mixed complete model h can be updated by $h_{t+1}=h_t-\eta g_{h,t}$ in the (t+1)-th round, and following Assumption 1, we can obtain

$$\begin{aligned}\mathcal{L}_{tE+1} &\leq \mathcal{L}_{tE+0}+\langle\nabla\mathcal{L}_{tE+0},(h_{tE+1}-h_{tE+0})\rangle\\ &\quad+\frac{L_1}{2}\|h_{tE+1}-h_{tE+0}\|_2^2\\ &=\mathcal{L}_{tE+0}-\eta\langle\nabla\mathcal{L}_{tE+0},g_{h,tE+0}\rangle\\ &\quad+\frac{L_1\eta^2}{2}\|g_{h,tE+0}\|_2^2.\end{aligned} \tag{19}$$

Taking the expectation of both sides of the inequality concerning the random variable ξ_{tE+0}, we obtain

$$\begin{aligned}\mathbb{E}[\mathcal{L}_{tE+1}] &\leq \mathcal{L}_{tE+0}-\eta\mathbb{E}[\langle\nabla\mathcal{L}_{tE+0},g_{h,tE+0}\rangle]\\ &\quad+\frac{L_1\eta^2}{2}\mathbb{E}[\|g_{h,tE+0}\|_2^2]\\ &\overset{(a)}{=}\mathcal{L}_{tE+0}-\eta\|\nabla\mathcal{L}_{tE+0}\|_2^2\\ &\quad+\frac{L_1\eta^2}{2}\mathbb{E}[\|g_{h,tE+0}\|_2^2]\\ &\overset{(b)}{\leq}\mathcal{L}_{tE+0}-\eta\|\nabla\mathcal{L}_{tE+0}\|_2^2\\ &\quad+\frac{L_1\eta^2}{2}(\mathbb{E}[\|g_{h,tE+0}\|]_2^2+\mathrm{Var}(g_{h,tE+0}))\\ &\overset{(c)}{=}\mathcal{L}_{tE+0}-\eta\|\nabla\mathcal{L}_{tE+0}\|_2^2\\ &\quad+\frac{L_1\eta^2}{2}(\|\nabla\mathcal{L}_{tE+0}\|_2^2+\mathrm{Var}(g_{h,tE+0}))\\ &\overset{(d)}{\leq}\mathcal{L}_{tE+0}-\eta\|\nabla\mathcal{L}_{tE+0}\|_2^2\\ &\quad+\frac{L_1\eta^2}{2}(\|\nabla\mathcal{L}_{tE+0}\|_2^2+\sigma^2)\\ &=\mathcal{L}_{tE+0}+(\frac{L_1\eta^2}{2}-\eta)\|\nabla\mathcal{L}_{tE+0}\|_2^2\\ &\quad+\frac{L_1\eta^2\sigma^2}{2}.\end{aligned} \tag{20}$$

(a), (c), (d) follow Assumption 2 and (b) follows $Var(x)=\mathbb{E}[x^2]-(\mathbb{E}[x]^2)$.

Taking the expectation of both sides of the inequality for the model h over E iterations, we obtain

$$\begin{aligned}\mathbb{E}[\mathcal{L}_{tE+1}] \leq \mathcal{L}_{tE+0}+(\frac{L_1\eta^2}{2}-\eta)\sum_{e=1}^{E}\|\nabla\mathcal{L}_{tE+e}\|_2^2\\ +\frac{L_1E\eta^2\sigma^2}{2}.\end{aligned} \tag{21}$$

Lemma 2: **Model Aggregation.** Given Assumptions 2 and 3, after the $(t+1)$-th local training round, the loss of any client before and after aggregating the small homogeneous feature extractors at the FL server is bounded by:

$$\mathbb{E}[\mathcal{L}_{(t+1)E+0}]\leq\mathbb{E}[\mathcal{L}_{tE+1}]+\eta\delta^2. \tag{22}$$

Proof 2:

$$\begin{aligned}\mathcal{L}_{(t+1)E+0}&=\mathcal{L}_{(t+1)E}+\mathcal{L}_{(t+1)E+0}-\mathcal{L}_{(t+1)E}\\ &\overset{(a)}{\approx}\mathcal{L}_{(t+1)E}+\eta\|\theta_{(t+1)E+0}-\theta_{(t+1)E}\|_2^2\\ &\overset{(b)}{\leq}\mathcal{L}_{(t+1)E}+\eta\delta^2.\end{aligned} \tag{23}$$

(a): we can use the gradient of parameter variations to approximate the loss variations, *i.e.*, $\Delta\mathcal{L} \approx \eta \cdot \|\Delta\theta\|_2^2$. (b) follows Assumption 3.

Taking the expectation of both sides of the inequality to the random variable ξ, we obtain

$$\mathbb{E}[\mathcal{L}_{(t+1)E+0}] \leq \mathbb{E}[\mathcal{L}_{tE+1}] + \eta\delta^2. \tag{24}$$

Theorem 1: **One Complete Round of FL**. Based on Lemma 1 and Lemma 2, for any client, after local training, model aggregation and receiving the new global homogeneous feature extractor, we have:

$$\begin{aligned}\mathbb{E}[\mathcal{L}_{(t+1)E+0}] \leq \mathcal{L}_{tE+0} + (\frac{L_1\eta^2}{2} - \eta)\sum_{e=0}^{E}\|\nabla\mathcal{L}_{tE+e}\|_2^2 \\ + \frac{L_1E\eta^2\sigma^2}{2} + \eta\delta^2.\end{aligned} \tag{25}$$

Proof 3: Substituting Lemma 1 into the right side of Lemma 2's inequality, we obtain

$$\begin{aligned}\mathbb{E}[\mathcal{L}_{(t+1)E+0}] \leq \mathcal{L}_{tE+0} + (\frac{L_1\eta^2}{2} - \eta)\sum_{e=0}^{E}\|\nabla\mathcal{L}_{tE+e}\|_2^2 \\ + \frac{L_1E\eta^2\sigma^2}{2} + \eta\delta^2.\end{aligned} \tag{26}$$

Theorem 2: **Non-convex Convergence Rate of pFedAFM.** With Theorem 1, for any client and an arbitrary constant $\epsilon > 0$, the following holds:

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{e=0}^{E-1}\|\nabla\mathcal{L}_{tE+e}\|_2^2 \leq \frac{\frac{1}{T}\sum_{t=0}^{T-1}[\mathcal{L}_{tE+0} - \mathbb{E}[\mathcal{L}_{(t+1)E+0}]]}{\eta - \frac{L_1\eta^2}{2}} \\ + \frac{\frac{L_1E\eta^2\sigma^2}{2} + \eta\delta^2}{\eta - \frac{L_1\eta^2}{2}} < \epsilon, \\ s.t.\ \eta < \frac{2(\epsilon - \delta^2)}{L_1(\epsilon + E\sigma^2)}.\end{aligned} \tag{27}$$

Therefore, we conclude that any client's local model can converge at a non-convex rate of $\epsilon \sim \mathcal{O}(1/T)$ in `pFedAFM` if the learning rates of the homogeneous feature extractor, local heterogeneous model and the trainable weight vector satisfy the above condition.

Proof 4: Interchanging the left and right sides of Eq. (26), we obtain

$$\begin{aligned}\sum_{e=0}^{E}\|\nabla\mathcal{L}_{tE+e}\|_2^2 \leq \frac{\mathcal{L}_{tE+0} - \mathbb{E}[\mathcal{L}_{(t+1)E+0}]}{\eta - \frac{L_1\eta^2}{2}} \\ + \frac{\frac{L_1E\eta^2\sigma^2}{2} + \eta\delta^2}{\eta - \frac{L_1\eta^2}{2}}.\end{aligned} \tag{28}$$

Taking the expectation of both sides of the inequality over rounds $t = [0, T-1]$ to W, we obtain

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{e=0}^{E-1}\|\nabla\mathcal{L}_{tE+e}\|_2^2 \leq \frac{\frac{1}{T}\sum_{t=0}^{T-1}[\mathcal{L}_{tE+0} - \mathbb{E}[\mathcal{L}_{(t+1)E+0}]]}{\eta - \frac{L_1\eta^2}{2}} \\ + \frac{\frac{L_1E\eta^2\sigma^2}{2} + \eta\delta^2}{\eta - \frac{L_1\eta^2}{2}}.\end{aligned} \tag{29}$$

Let $\Delta = \mathcal{L}_{t=0} - \mathcal{L}^* > 0$, then $\sum_{t=0}^{T-1}[\mathcal{L}_{tE+0} - \mathbb{E}[\mathcal{L}_{(t+1)E+0}]] \leq \Delta$, we can get

$$\frac{1}{T}\sum_{t=0}^{T-1}\sum_{e=0}^{E-1}\|\nabla\mathcal{L}_{tE+e}\|_2^2 \leq \frac{\frac{\Delta}{T} + \frac{L_1E\eta^2\sigma^2}{2} + \eta\delta^2}{\eta - \frac{L_1\eta^2}{2}}. \tag{30}$$

If the above equation converges to a constant ϵ, *i.e.*,

$$\frac{\frac{\Delta}{T} + \frac{L_1E\eta^2\sigma^2}{2} + \eta\delta^2}{\eta - \frac{L_1\eta^2}{2}} < \epsilon, \tag{31}$$

then

$$T > \frac{\Delta}{\epsilon(\eta - \frac{L_1\eta^2}{2}) - \frac{L_1E\eta^2\sigma^2}{2} - \eta\delta^2}. \tag{32}$$

Since $T > 0, \Delta > 0$, we can get

$$\epsilon(\eta - \frac{L_1\eta^2}{2}) - \frac{L_1E\eta^2\sigma^2}{2} - \eta\delta^2 > 0. \tag{33}$$

Solving the above inequality yields

$$\eta < \frac{2(\epsilon - \delta^2)}{L_1(\epsilon + E\sigma^2)}. \tag{34}$$

Since ϵ, L_1, σ^2, δ^2 are all constants greater than 0, η has solutions. Therefore, when the learning rate η satisfies the above condition, any client's local mixed complete heterogeneous model can converge. Notice that the learning rate of the local complete heterogeneous model involves $\{\eta_\theta, \eta_\omega, \eta_{\boldsymbol{\alpha}}\}$, so it's crucial to set reasonable them to ensure model convergence. Since all terms on the right side of Eq. (30) except for $1/T$ are constants, hence `pFedAFM`'s non-convex convergence rate is $\epsilon \sim \mathcal{O}(1/T)$.