A APPENDIX

A.1 CONVERGENCE ANALYSIS

Herein, we provide the proofs of the lemmas and theorem shown in Section 5.

A.1.1 PRELIMINARIES

FedLAMA periodically chooses a few layers that will be less frequently synchronized. We call these layers Least Critical Layers (LCL) for short.

Notations – All vectors in this paper are column vectors. $\mathbf{x} \in \mathbb{R}^d$ denotes the parameters of one local model and m is the number of workers. The stochastic gradient computed from a single training data point $\boldsymbol{\xi}$ is denoted by $g(\mathbf{x}, \boldsymbol{\xi})$. For convenience, we use $g(\mathbf{x})$ instead. The full batch gradient is denoted by $\nabla F(\mathbf{x})$. We use $\|\cdot\|$ and $\|\cdot\|_{op}$ to denote $l2$ norm and matrix operator norm, respectively.

Objective Function – In this paper, we consider federated optimization problems as follows.

$$\min_{\mathbf{x}\in\mathbb{R}^d} \left[F(\mathbf{x}) := \sum_{i=1}^{m} p_i F_i(\mathbf{x}) \right], \tag{10}$$

where $p_i = n_i/n$ is the ratio of local data to the total dataset, and $F_i(\mathbf{x}) = \frac{1}{n_i}\sum_{\xi\in\mathcal{D}} f_i(\mathbf{x}, \xi)$ is the local objective function of client i. n is the global dataset size and n_i is the local dataset size. Note that, by definition, $\sum_{i=1}^{m} p_i = 1$.

Averaging Matrix – We define a time-varying averaging matrix $\mathbf{W}_k \in \mathbb{R}^{dm\times dm}$ as follows.

$$\mathbf{W_k} = \begin{cases} \mathbf{P}, & \text{if } k \bmod \tau_{min} \text{ is } 0 \\ \mathbf{J}, & \text{if } k \bmod \tau_{max} \text{ is } 0 \\ \mathbf{I}, & \text{otherwise} \end{cases} \tag{11}$$

$\mathbf{I}$ is an identity matrix, $\mathbf{P}$ is also a time-varying averaging matrix, and $\mathbf{J}$ is a full averaging matrix. First, $\mathbf{P}_i^1$ is a $d \times d$ diagonal matrix that has 1 for the diagonal elements that correspond to the LCL parameters and p_i for all the other diagonal elements. Likewise, $\mathbf{P}_i^0$ is another $d \times d$ diagonal matrix that has 0 for the diagonal elements that correspond to the LCL parameters and p_i for all the other diagonal elements. Then, $\mathbf{P}$ is defined as follows.

$$\mathbf{P} = \begin{cases} \mathbf{P}^1, \text{ for } m \text{ diagonal blocks} \\ \mathbf{P}^0, \text{ for all the other blocks} \end{cases} \tag{12}$$

The i^{th} block column of $\mathbf{P}$ consists of $\mathbf{P}_i^1$ and $\mathbf{P}_i^0$ following the above definition.

Here we present an example of $\mathbf{P}$ where $m = 2$ and $d = 2$. In this example, $p_0 = 1/3$ and $p_1 = 2/3$. Saying the LCL is the second parameter, $\mathbf{P}$ is defined as follows.

$$\mathbf{P}_0^1 = \begin{bmatrix} \frac{1}{3} & 0 \\ 0 & 1 \end{bmatrix}, \mathbf{P}_0^0 = \begin{bmatrix} \frac{1}{3} & 0 \\ 0 & 0 \end{bmatrix}, \mathbf{P}_1^1 = \begin{bmatrix} \frac{2}{3} & 0 \\ 0 & 1 \end{bmatrix}, \mathbf{P}_1^0 = \begin{bmatrix} \frac{2}{3} & 0 \\ 0 & 0 \end{bmatrix} \tag{13}$$

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_0^1 & \mathbf{P}_1^0 \\ \mathbf{P}_0^0 & \mathbf{P}_1^1 \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & 0 & \frac{2}{3} & 0 \\ 0 & 1 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{2}{3} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{14}$$

The full-averaging matrix $\mathbf{J}$ is defined as follows. First, $\mathbf{J}_i$ is a $d \times d$ diagonal matrix that has p_i for the diagonal elements. Then, $\mathbf{J}$ consists of $m \times m$ blocks of $\mathbf{J}_i$ such that each column block is m of $\mathbf{J}_i$ blocks. Here we present an example of $\mathbf{J}$ where $m = 2$ and $d = 2$ as follows.

$$\mathbf{J}_0 = \begin{bmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{3} \end{bmatrix}, \mathbf{J}_1 = \begin{bmatrix} \frac{2}{3} & 0 \\ 0 & \frac{2}{3} \end{bmatrix} \tag{15}$$

$$\mathbf{J} = \begin{bmatrix} \mathbf{J}_0 & \mathbf{J}_1 \\ \mathbf{J}_0 & \mathbf{J}_1 \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & 0 & \frac{2}{3} & 0 \\ 0 & \frac{1}{3} & 0 & \frac{2}{3} \\ \frac{1}{3} & 0 & \frac{2}{3} & 0 \\ 0 & \frac{1}{3} & 0 & \frac{2}{3} \end{bmatrix}. \tag{16}$$

The averaging matrix $\mathbf{P}$ and $\mathbf{J}$ have the following properties:

1. $\mathbf{P}\mathbf{1}_{dm} = \mathbf{1}_{dm}, \mathbf{J}\mathbf{1}_{dm} = \mathbf{1}_{dm}$.
2. The product of any two averaging matrices consists only of diagonal block matrices because all the blocks in $\mathbf{P}$ and $\mathbf{J}$ are diagonal.
3. $\mathbf{PJ} = \mathbf{JP} = \mathbf{J}$ regardless of which layers are chosen as the LCL.
4. $\mathbf{PP} = \mathbf{P}$ regardless of which layers are chosen as the LCL.

Vectorization – We define a vectorized form of m local model parameters $\mathbf{x}_k \in \mathbb{R}^{dm}$, its stochastic gradients $\mathbf{g}_k \in \mathbb{R}^{dm}$, and the full gradients $\mathbf{f}_k \in \mathbb{R}^{dm}$ as follows

$$\begin{aligned} \mathbf{x}_k &= vec\left\{\mathbf{x}_k^1, \mathbf{x}_k^2, \cdots, \mathbf{x}_k^m\right\} \\ \mathbf{g}_k &= vec\left\{g_1(\mathbf{x}_k^1), g_2(\mathbf{x}_k^2), \cdots, g_m(\mathbf{x}_k^m)\right\} \\ \mathbf{f}_k &= vec\left\{\nabla F_1(\mathbf{x}_k^1), \nabla F_2(\mathbf{x}_k^2), \cdots, \nabla F_m(\mathbf{x}_k^m)\right\}. \end{aligned} \tag{17}$$

The full model aggregation can be written using the vectorized form of local models $\mathbf{x}_k$ and the averaging matrix $\mathbf{J}$ as follows.

$$\mathbf{J}\mathbf{x}_k = \begin{bmatrix} \frac{1}{3} & 0 & \frac{2}{3} & 0 \\ 0 & \frac{1}{3} & 0 & \frac{2}{3} \\ \frac{1}{3} & 0 & \frac{2}{3} & 0 \\ 0 & \frac{1}{3} & 0 & \frac{2}{3} \end{bmatrix} \begin{bmatrix} x_k^{(1,1)} \\ x_k^{(1,2)} \\ x_k^{(2,1)} \\ x_k^{(2,2)} \end{bmatrix} = \begin{bmatrix} (x_k^{(1,1)} + 2x_k^{(2,1)})/3 \\ (x_k^{(1,2)} + 2x_k^{(2,2)})/3 \\ (x_k^{(1,1)} + 2x_k^{(2,1)})/3 \\ (x_k^{(1,2)} + 2x_k^{(2,2)})/3 \end{bmatrix} \tag{18}$$

where $x_k^{(i,j)}$ is the j^{th} model parameter of local model i at iteration k.

We also define the following additional vectorized forms of the weighted model parameters and gradients for convenience.

$$\begin{aligned} \hat{\mathbf{x}}_k &= \text{vec}\left\{\sqrt{p_1}\mathbf{x}_k^1, \sqrt{p_2}\mathbf{x}_k^2, \cdots, \sqrt{p_m}\mathbf{x}_k^m\right\} \\ \hat{\mathbf{g}}_k &= \text{vec}\left\{\sqrt{p_1}g_1(\mathbf{x}_k^1), \sqrt{p_2}g_2(\mathbf{x}_k^2), \cdots, \sqrt{p_m}g_m(\mathbf{x}_k^m)\right\} \\ \hat{\mathbf{f}}_k &= \text{vec}\left\{\sqrt{p_1}\nabla F_1(\mathbf{x}_k^1), \sqrt{p_2}\nabla F_2(\mathbf{x}_k^2), \cdots, \sqrt{p_m}\nabla F_m(\mathbf{x}_k^m)\right\} \end{aligned} \tag{19}$$

Assumptions – We analyze the convergence rate of FedLAMA under the following assumptions.

1. (Smoothness). Each local objective function is L-smooth, that is, $\|\nabla F_i(\mathbf{x}) - \nabla F_i(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|, \forall i \in \{1, \cdots, m\}$.
2. (Unbiased Gradient). The stochastic gradient at each client is an unbiased estimator of the local full-batch gradient: $\mathbb{E}_\xi[g_i(\mathbf{x}, \xi)] = \nabla F_i(\mathbf{x})$.
3. (Bounded Variance). The stochastic gradient at each client has bounded variance: $\mathbb{E}_\xi[\|g_i(\mathbf{x}, \xi) - \nabla F_i(\mathbf{x})\|^2 \leq \sigma^2], \forall i \in \{1, \cdots, m\}, \sigma^2 \geq 0$.
4. (Bounded Dissimilarity). For any sets of weights $\{p_i \geq 0\}_{i=1}^m, \sum_{i=1}^m p_i = 1$, there exist constants $\beta^2 \geq 1$ and $\kappa^2 \geq 0$ such that $\sum_{i=1}^m p_i\|\nabla F_i(\mathbf{x})\|^2 \leq \beta^2\|\sum_{i=1}^m p_i \nabla F_i(\mathbf{x})\|^2 + \kappa^2$. If local objective functions are identical to each other, $\beta^2 = 1$ and $\kappa^2 = 0$.

A.1.2 PROOFS

Theorem 5.1. *Suppose all m local models are initialized to the same point $\mathbf{u}_1$. Under Assumption $1 \sim 4$, if FedLAMA runs for K iterations and the learning rate satisfies $\eta \leq \min\left\{\frac{1}{2(\tau_{max}-1)L}, \frac{1}{L\sqrt{2\tau_{max}(\tau_{max}-1)(2\beta^2+1)}}\right\}$, FedLAMA ensures*

$$\mathbb{E}\left[\frac{1}{K}\sum_{i=1}^{K}\|\nabla F(\mathbf{u}_k)\|^2\right] \leq \frac{4}{\eta K}\left(\mathbb{E}\left[F(\mathbf{u}_1)-F(\mathbf{u}_*)\right]\right) + 4\eta\sum_{i=1}^{m}p_i^2 L\sigma^2 + 3\eta^2(\tau_{max}-1)L^2\sigma^2 + 6\eta^2\tau_{max}(\tau_{max}-1)L^2\kappa^2, \tag{20}$$

where $\mathbf{u}_*$ indicates a local minimum.

Proof. Based on Lemma 5.1 and 5.2, we have

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla F(\mathbf{u}_k)\|^2\right] \leq \frac{2}{\eta K}\left(\mathbb{E}\left[F(\mathbf{u}_1)-F(\mathbf{u}_*)\right]\right) + 2\eta\sum_{i=1}^{m}p_i^2 L\sigma^2 + L^2\left(\frac{\eta^2(\tau_{max}-1)\sigma_j^2}{1-A} + \frac{A\beta^2}{KL^2(1-A)}\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla F(\mathbf{u}_k)\|^2\right] + \frac{A\kappa^2}{L^2(1-A)}\right).$$

After re-writing the left-hand side and a minor rearrangement, we have

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla F(\mathbf{u}_k)\|^2\right] \leq \frac{2}{\eta K}\left(\mathbb{E}\left[F(\mathbf{u}_1)-F(\mathbf{u}_*)\right]\right) + 2\eta\sum_{i=1}^{m}p_i^2 L\sigma^2 + \frac{1}{K}\sum_{k=1}^{K}\frac{A\beta^2}{1-A}\mathbb{E}\left[\|\nabla F(\mathbf{u}_k)\|^2\right] + L^2\left(\frac{\eta^2(\tau_{max}-1)\sigma^2}{1-A} + \frac{A\kappa^2}{L^2(1-A)}\right).$$

By moving the third term on the right-hand side to the left-hand side, we have

$$\frac{1}{K}\sum_{k=1}^{K}\left(1-\frac{A\beta^2}{1-A}\right)\mathbb{E}\left[\|\nabla_j F(\mathbf{u}_k)\|^2\right] \leq \frac{2}{\eta K}\left(\mathbb{E}\left[F(\mathbf{u}_1)-F(\mathbf{u}_*)\right]\right) + 2\eta\sum_{i=1}^{m}p_i^2 L\sigma^2 + L^2\left(\frac{\eta^2(\tau_{max}-1)\sigma^2}{1-A} + \frac{A\kappa^2}{L^2(1-A)}\right). \tag{21}$$

If $A \leq \frac{1}{2\beta^2+1}$, then $\frac{A\beta^2}{1-A} \leq \frac{1}{2}$. Therefore, (21) can be simplified as follows.

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla F(\mathbf{u}_k)\|^2\right] \leq \frac{4}{\eta K}\left(\mathbb{E}\left[F(\mathbf{u}_1)-F(\mathbf{u}_*)\right]\right) + 4\eta\sum_{i=1}^{m}p_i^2 L\sigma^2 + 2L^2\left(\frac{\eta^2(\tau_{max}-1)\sigma^2}{1-A}\right) + 2\frac{A\kappa^2}{1-A}. \tag{22}$$

The learning rate condition $A \leq \frac{1}{2\beta^2+1}$ also ensures that $\frac{1}{1-A} \leq 1 + \frac{1}{2\beta^2}$. Based on Assumption 4, $\frac{1}{2\beta^2} \leq \frac{2}{3}$, and thus $\frac{1}{1-A} \leq \frac{2}{3}$. Therefore, we have

$$\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla F(\mathbf{u}_k)\|^2\right] \leq \frac{4}{\eta K}\left(\mathbb{E}\left[F(\mathbf{u}_1)-F(\mathbf{u}_*)\right]\right) + 4\eta\sum_{i=1}^{m}p_i^2 L\sigma^2 + 3\eta^2(\tau_{max}-1)L^2\sigma^2 + 6\eta^2\tau_{max}(\tau_{max}-1)L^2\kappa^2.$$

We complete the proof. □

Learning Rate Constraints – In Theorem 5.3, we have two learning rate constraints, one from (22) and the other from (51) as follows.

$$A < \frac{1}{2\beta^2 + 1} \qquad \text{from (22)}$$
$$A < 1 \qquad \text{from (51)}$$

After a minor rearrangement, we have a unified learning rate constraint as follows.

$$\eta \leq \min\left\{ \frac{1}{2(\tau_{max}-1)L}, \frac{1}{L\sqrt{2\tau_{max}(\tau_{max}-1)(2\beta^2+1)}} \right\}$$

Lemma 5.1. *(Framework) Under Assumption* $1 \sim 3$*, if the learning rate satisfies* $\eta \leq \frac{1}{2L}$*, FedLAMA ensures*

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla F(\mathbf{u}_k)\right\|^2\right] \leq \frac{2}{\eta K}\mathbb{E}\left[F(\mathbf{u}_1) - F(\mathbf{u}_*)\right] + 2\eta L\sigma^2\sum_{i=1}^{m}(p_i)^2 \\ + \frac{L^2}{K}\sum_{k=1}^{K}\sum_{i=1}^{m} p_i\,\mathbb{E}\left[\left\|\mathbf{u}_k - \mathbf{x}_k^i\right\|^2\right].\end{aligned} \tag{23}$$

Proof. Based on Assumption 1, we have

$$\mathbb{E}\left[F(\mathbf{u}_{k+1}) - F(\mathbf{u}_k)\right] \leq -\eta\underbrace{\mathbb{E}\left[\left\langle\nabla F(\mathbf{u}_k), \sum_{i=1}^{m} p_i g_i(\mathbf{x}_k^i)\right\rangle\right]}_{T_1} + \frac{\eta^2 L}{2}\underbrace{\mathbb{E}\left[\left\|\sum_{i=1}^{m} p_i g_i(\mathbf{x}_k^i)\right\|^2\right]}_{T_2} \tag{24}$$

First, T_1 can be rewritten as follows.

$$\begin{aligned}T_1 &= \mathbb{E}\left[\left\langle\nabla F(\mathbf{u}_k), \sum_{i=1}^{m} p_i\left(g_i(\mathbf{x}_k^i) - \nabla F_i(\mathbf{x}_k^i)\right)\right\rangle\right] + \mathbb{E}\left[\left\langle\nabla F(\mathbf{u}_k), \sum_{i=1}^{m} p_i\nabla F_i(\mathbf{x}_k^i)\right\rangle\right] \\ &= \mathbb{E}\left[\left\langle\nabla F(\mathbf{u}_k), \sum_{i=1}^{m} p_i\nabla F_i(\mathbf{x}_k^i)\right\rangle\right] \\ &= \frac{1}{2}\left\|\nabla F(\mathbf{u}_k)\right\|^2 + \frac{1}{2}\mathbb{E}\left[\left\|\sum_{i=1}^{m} p_i\nabla F_i(\mathbf{x}_k^i)\right\|^2\right] - \frac{1}{2}\mathbb{E}\left[\left\|\nabla F(\mathbf{u}_k) - \sum_{i=1}^{m} p_i\nabla F_i(\mathbf{x}_k^i)\right\|^2\right],\end{aligned} \tag{25}$$

where the last equality holds based on a basic equality: $2\mathbf{a}^\top\mathbf{b} = \|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - \|\mathbf{a}-\mathbf{b}\|^2$.

Then, T_2 can be bounded as follows.

$$\begin{aligned}T_2 &= \mathbb{E}\left[\left\|\sum_{i=1}^{m} p_i\left(g_i(\mathbf{x}_k^i) - \mathbb{E}\left[g_i(\mathbf{x}_k^i)\right]\right) + \sum_{i=1}^{m} p_i\,\mathbb{E}\left[g_i(\mathbf{x}_k^i)\right]\right\|^2\right] \\ &= \mathbb{E}\left[\left\|\sum_{i=1}^{m} p_i\left(g_i(\mathbf{x}_k^i) - \nabla F_i(\mathbf{x}_k^i)\right) + \sum_{i=1}^{m} p_i\nabla F_i(\mathbf{x}_k^i)\right\|^2\right] \\ &\leq 2\,\mathbb{E}\left[\left\|\sum_{i=1}^{m} p_i\left(g_i(\mathbf{x}_k^i) - \nabla F_i(\mathbf{x}_k^i)\right)\right\|^2\right] + 2\,\mathbb{E}\left[\left\|\sum_{i=1}^{m} p_i\nabla F_i(\mathbf{x}_k^i)\right\|^2\right] \\ &= 2\sum_{i=1}^{m} p_i^2\,\mathbb{E}\left[\left\|g_i(\mathbf{x}_k^i) - \nabla F_i(\mathbf{x}_k^i)\right\|^2\right] + 2\,\mathbb{E}\left[\left\|\sum_{i=1}^{m} p_i\nabla F_i(\mathbf{x}_k^i)\right\|^2\right] \\ &\leq 2\sigma^2\sum_{i=1}^{m} p_i^2 + 2\,\mathbb{E}\left[\left\|\sum_{i=1}^{m} p_i\nabla F_i(\mathbf{x}_k^i)\right\|^2\right],\end{aligned} \tag{26}$$

where the last equality holds because $g_i(\mathbf{x}_k^i) - \nabla F_i(\mathbf{x}_k^i)$ has 0 mean and is independent across i, and the last inequality follows Assumption 3.

By plugging in (25) and (26) into (24), we have the following.

$$\begin{aligned}
\mathbb{E}\left[F(\mathbf{u}_{k+1}) - F(\mathbf{u}_k)\right] \leq & -\frac{\eta}{2}\left\|\nabla F(\mathbf{u}_k)\right\|^2 - \frac{\eta}{2}\,\mathbb{E}\left[\left\|\sum_{i=1}^m p_i \nabla F_i(\mathbf{x}_k^i)\right\|^2\right] \\
& + \frac{\eta}{2}\,\mathbb{E}\left[\left\|\nabla F(\mathbf{u}_k) - \sum_{i=1}^m p_i \nabla F_i(\mathbf{x}_k^i)\right\|^2\right] + \eta^2 L\sigma^2 \sum_{i=1}^m p_i^2 \\
& + \eta^2 L\,\mathbb{E}\left[\left\|\sum_{i=1}^m p_i \nabla F_i(\mathbf{x}_k^i)\right\|^2\right] \\
= & -\frac{\eta}{2}\left\|\nabla F(\mathbf{u}_k)\right\|^2 - \frac{\eta}{2}(1-2\eta L)\,\mathbb{E}\left[\left\|\sum_{i=1}^m p_i \nabla F_i(\mathbf{x}_k^i)\right\|^2\right] \\
& + \frac{\eta}{2}\,\mathbb{E}\left[\left\|\nabla F(\mathbf{u}_k) - \sum_{i=1}^m p_i \nabla F_i(\mathbf{x}_k^i)\right\|^2\right] + \eta^2 L\sigma^2 \sum_{i=1}^m p_i^2
\end{aligned}$$

If $\eta \leq \frac{1}{2L}$, it follows

$$\begin{aligned}
\frac{\mathbb{E}\left[F(\mathbf{u}_{k+1}) - F(\mathbf{u}_k)\right]}{\eta} \leq & -\frac{1}{2}\left\|\nabla F(\mathbf{u}_k)\right\|^2 + \eta L\sigma^2 \sum_{i=1}^m p_i^2 \\
& + \frac{1}{2}\,\mathbb{E}\left[\left\|\nabla F(\mathbf{u}_k) - \sum_{i=1}^m p_i \nabla F_i(\mathbf{x}_k^i)\right\|^2\right] \\
\leq & -\frac{1}{2}\left\|\nabla F(\mathbf{u}_k)\right\|^2 + \eta L\sigma^2 \sum_{i=1}^m p_i^2 \qquad (27) \\
& + \frac{1}{2}\sum_{i=1}^m p_i\,\mathbb{E}\left[\left\|\nabla F_i(\mathbf{u}_k) - \nabla F_i(\mathbf{x}_k^i)\right\|^2\right] \\
\leq & -\frac{1}{2}\left\|\nabla F(\mathbf{u}_k)\right\|^2 + \eta L\sigma^2 \sum_{i=1}^m p_i^2 + \frac{L^2}{2}\sum_{i=1}^m p_i\,\mathbb{E}\left[\left\|\mathbf{u}_k - \mathbf{x}_k^i\right\|^2\right],
\end{aligned}$$

where (27) holds based on the convexity of $\ell 2$ norm and Jensen's inequality.

By taking expectation and averaging across K iterations, we have.

$$\begin{aligned}
\frac{1}{K}\sum_{k=1}^K \frac{\mathbb{E}\left[F(\mathbf{u}_{k+1}) - F(\mathbf{u}_k)\right]}{\eta} \leq & -\frac{1}{2K}\sum_{k=1}^K \left\|\nabla F(\mathbf{u}_k)\right\|^2 + \eta L\sigma^2 \sum_{i=1}^m p_i^2 \\
& + \frac{L^2}{2K}\sum_{k-1}^K \sum_{i=1}^m p_i\,\mathbb{E}\left[\left\|\mathbf{u}_k - \mathbf{x}_k^i\right\|^2\right].
\end{aligned}$$

After a minor rearrangement, we have a telescoping sum as follows.

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla F(\mathbf{u}_k)\|^2\right] &\leq \frac{2}{\eta K}\mathbb{E}\left[F(\mathbf{u}_1)-F(\mathbf{u}_{k+1})\right]+2\eta L\sigma^2\sum_{i=1}^{m}p_i^2\\ &\quad+\frac{L^2}{K}\sum_{k=1}^{K}\sum_{i=1}^{m}p_i\,\mathbb{E}\left[\left\|\mathbf{u}_k-\mathbf{x}_k^i\right\|^2\right]\\ &\leq \frac{2}{\eta K}\mathbb{E}\left[F(\mathbf{u}_1)-F(\mathbf{u}_*)\right]+2\eta L\sigma^2\sum_{i=1}^{m}p_i^2\\ &\quad+\frac{L^2}{K}\sum_{k=1}^{K}\sum_{i=1}^{m}p_i\,\mathbb{E}\left[\left\|\mathbf{u}_k-\mathbf{x}_k^i\right\|^2\right],\end{aligned}$$

where $\mathbf{u}_*$ indicates the local minimum. Here, we complete the proof. □

Lemma 5.2. *(Model Discrepancy) Under Assumption* $1 \sim 4$*, if the learning rate satisfies* $\eta < \frac{1}{2(\tau_{max}-1)L}$*, FedLAMA ensures*

$$\begin{aligned}\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{m}p_i\,\mathbb{E}\left[\left\|\mathbf{u}_k-\mathbf{x}_k^i\right\|^2\right] &\leq \frac{2\eta^2(\tau_{max}-1)\sigma^2}{1-A}+\frac{A\kappa^2}{L^2(1-A)}\\ &\quad+\frac{A\beta^2}{KL^2(1-A)}\sum_{k=1}^{K}\mathbb{E}\left[\|\nabla F(\mathbf{u}_k)\|^2\right],\end{aligned}\tag{28}$$

where $A=4\eta^2(\tau_{max}-1)^2L^2$ *and* τ_{max} *is the largest averaging interval across all the layers.*

Proof. We begin with rewriting the weighted average of the squared distance using the vectorized form of the local models as follows.

$$\begin{aligned}\sum_{i=1}^{m}p_i\left\|\mathbf{u}_k-\mathbf{x}_k^i\right\|^2 &= \sum_{i=1}^{m}\left\|\sqrt{p_i}\left(\mathbf{u}_k-\mathbf{x}_k^i\right)\right\|^2\\ &= \|\mathbf{J}\hat{\mathbf{x}}_k-\hat{\mathbf{x}}_k\|^2\\ &= \|(\mathbf{J}-\mathbf{I})\hat{\mathbf{x}}_k\|^2,\end{aligned}\tag{29}$$

where (29) holds by the commutative property of multiplication.

Then, according to the parameter update rule, we have

$$\begin{aligned}(\mathbf{J}-\mathbf{I})\hat{\mathbf{x}}_k &= (\mathbf{J}-\mathbf{I})\mathbf{W}_{k-1}(\hat{\mathbf{x}}_{k-1}-\eta\hat{\mathbf{g}}_{k-1})\\ &= (\mathbf{J}-\mathbf{I})\mathbf{W}_{k-1}\hat{\mathbf{x}}_{k-1}-(\mathbf{J}-\mathbf{W}_{k-1})\eta\hat{\mathbf{g}}_{k-1},\end{aligned}\tag{30}$$

where (30) holds because $\mathbf{JW}=\mathbf{J}$ based on the averaging matrix property 3, and $\mathbf{IW}=\mathbf{W}$.

Then, expanding the expression of $\mathbf{x}_{k-1}$, we have

$$\begin{aligned}(\mathbf{J}-\mathbf{I})\hat{\mathbf{x}}_k &= (\mathbf{J}-\mathbf{I})\mathbf{W}_{k-1}(\mathbf{W}_{k-2}(\hat{\mathbf{x}}_{k-2}-\eta\hat{\mathbf{g}}_{k-2}))-(\mathbf{J}-\mathbf{W}_{k-1})\eta\hat{\mathbf{g}}_{k-1}\\ &= (\mathbf{J}-\mathbf{I})\mathbf{W}_{k-1}\mathbf{W}_{k-2}\hat{\mathbf{x}}_{k-2}-(\mathbf{J}-\mathbf{W}_{k-1}\mathbf{W}_{k-2})\eta\hat{\mathbf{g}}_{k-2}-(\mathbf{J}-\mathbf{W}_{k-1})\eta\hat{\mathbf{g}}_{k-1}.\end{aligned}$$

Repeating the same procedure for $\hat{\mathbf{x}}_{k-2}, \hat{\mathbf{x}}_{k-3}, \cdots, \hat{\mathbf{x}}_2$, we have

$$\begin{aligned}(\mathbf{J}-\mathbf{I})\hat{\mathbf{x}}_k &= (\mathbf{J}-\mathbf{I})\prod_{s=1}^{k-1}\mathbf{W}_s\hat{\mathbf{x}}_1-\eta\sum_{s=1}^{k-1}(\mathbf{J}-\prod_{l=s}^{k-1}\mathbf{W}_l)\hat{\mathbf{g}}_s\\ &= -\eta\sum_{s=1}^{k-1}(\mathbf{J}-\prod_{l=s}^{k-1}\mathbf{W}_l)\hat{\mathbf{g}}_s,\end{aligned}\tag{31}$$

where (31) holds because $\mathbf{x}_1^i$ is the same across all the workers and thus $(\mathbf{J}-\mathbf{I})\hat{\mathbf{x}}_1=0$.

Based on (31), we have

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{m} p_i\,\mathbb{E}\left[\left\|\mathbf{u}_k-\mathbf{x}_k^i\right\|^2\right] \\
&\quad=\frac{1}{K}\sum_{k=1}^{K}\left(\mathbb{E}\left[\|(\mathbf{J}-\mathbf{I})\hat{\mathbf{x}}_k\|^2\right]\right) \\
&\quad=\frac{1}{K}\sum_{k=1}^{K}\left(\eta^2\,\mathbb{E}\left[\left\|\sum_{s=1}^{k-1}(\mathbf{J}-\prod_{l=s}^{k-1}\mathbf{W}_l)\hat{\mathbf{g}}_s\right\|^2\right]\right) \\
&\quad=\frac{1}{K}\sum_{k=1}^{K}\left(\eta^2\,\mathbb{E}\left[\left\|\sum_{s=1}^{k-1}(\mathbf{J}-\prod_{l=s}^{k-1}\mathbf{W}_l)(\hat{\mathbf{g}}_s-\hat{\mathbf{f}}_s)+\sum_{s=1}^{k-1}(\mathbf{J}-\prod_{l=s}^{k-1}\mathbf{W}_l)\hat{\mathbf{f}}_s\right\|^2\right]\right) \\
&\quad\le\frac{2\eta^2}{K}\left(\underbrace{\sum_{k=1}^{K}\mathbb{E}\left[\left\|\sum_{s=1}^{k-1}(\mathbf{J}-\prod_{l=s}^{k-1}\mathbf{W}_l)(\hat{\mathbf{g}}_s-\hat{\mathbf{f}}_s)\right\|^2\right]}_{T_3}+\underbrace{\sum_{k=1}^{K}\mathbb{E}\left[\left\|\sum_{s=1}^{k-1}(\mathbf{J}-\prod_{l=s}^{k-1}\mathbf{W}_l)\hat{\mathbf{f}}_s\right\|^2\right]}_{T_4}\right)
\end{aligned}
\tag{32}
$$

where (32) holds based on the convexity of $\ell 2$ norm and Jensen's inequality. Now, we focus on bounding T_3 and T_4, separately.

Bounding T_3

$$
\begin{aligned}
&\sum_{k=1}^{K}\mathbb{E}\left[\left\|\sum_{s=1}^{k-1}(\mathbf{J}-\prod_{l=s}^{k-1}\mathbf{W}_l)(\hat{\mathbf{g}}_s-\hat{\mathbf{f}}_s)\right\|^2\right] \\
&\qquad=\sum_{k=1}^{K}\sum_{s=1}^{k-1}\mathbb{E}\left[\left\|(\mathbf{J}-\prod_{l=s}^{k-1}\mathbf{W}_l)(\hat{\mathbf{g}}_s-\hat{\mathbf{f}}_s)\right\|^2\right] &(33)\\
&\qquad\le\sum_{k=1}^{K}\sum_{s=1}^{k-1}\mathbb{E}\left[\left\|(\hat{\mathbf{g}}_s-\hat{\mathbf{f}}_s)\right\|^2\left\|(\mathbf{J}-\prod_{l=s}^{k-1}\mathbf{W}_l)\right\|_{op}^2\right], &(34)
\end{aligned}
$$

where (33) holds because $\hat{\mathbf{g}}_s-\hat{\mathbf{f}}_s$ has 0 mean and independent across s, and (34) holds based on Lemma A.1.

Without loss of generality, we replace k with $a\tau_{max}+b$, where a is the communication round index and b is the iteration index within each communication round. Then, we have

$$
\begin{aligned}
&\sum_{a=0}^{K/\tau_{max}-1}\sum_{b=1}^{\tau_{max}}\sum_{s=1}^{a\tau_{max}+b-1}\mathbb{E}\left[\left\|(\hat{\mathbf{g}}_s-\hat{\mathbf{f}}_s)\right\|^2\left\|(\mathbf{J}-\prod_{l=s}^{k-1}\mathbf{W}_l)\right\|_{op}^2\right] \\
&=\sum_{a=0}^{K/\tau_{max}-1}\sum_{b=1}^{\tau_{max}}\sum_{s=1}^{a\tau}\mathbb{E}\left[\left\|(\hat{\mathbf{g}}_s-\hat{\mathbf{f}}_s)\right\|^2\left\|(\mathbf{J}-\prod_{l=s}^{a\tau_{max}+b-1}\mathbf{W}_l)\right\|_{op}^2\right] \\
&\quad+\sum_{a=0}^{K/\tau_{max}-1}\sum_{b=1}^{\tau_{max}}\sum_{s=a\tau_{max}+1}^{a\tau_{max}+b-1}\mathbb{E}\left[\left\|(\hat{\mathbf{g}}_s-\hat{\mathbf{f}}_s)\right\|^2\left\|(\mathbf{J}-\prod_{l=s}^{a\tau_{max}+b-1}\mathbf{W}_l)\right\|_{op}^2\right] \\
&=\sum_{a=0}^{K/\tau_{max}-1}\sum_{b=1}^{\tau_{max}}\sum_{s=a\tau_{max}+1}^{a\tau_{max}+b-1}\mathbb{E}\left[\left\|(\hat{\mathbf{g}}_s-\hat{\mathbf{f}}_s)\right\|^2\left\|(\mathbf{J}-\prod_{l=s}^{a\tau_{max}+b-1}\mathbf{W}_l)\right\|_{op}^2\right] \qquad (35)\\
&=\sum_{a=0}^{K/\tau_{max}-1}\sum_{b=1}^{\tau_{max}}\sum_{s=a\tau_{max}+1}^{a\tau+b-1}\mathbb{E}\left[\left\|(\hat{\mathbf{g}}_s-\hat{\mathbf{f}}_s)\right\|^2\right] \qquad (36)\\
&=\sum_{a=0}^{K/\tau_{max}-1}\sum_{b=1}^{\tau_{max}}\sum_{s=a\tau_{max}+1}^{a\tau_{max}+b-1}\sum_{i=1}^{m}p_i\,\mathbb{E}\left[\left\|(g_i(\mathbf{x}_s^i)-\nabla F_i(\mathbf{x}_s^i))\right\|^2\right] \\
&\leq\sum_{a=0}^{K/\tau_{max}-1}\sum_{b=1}^{\tau_{max}}\sum_{s=a\tau_{max}+1}^{a\tau_{max}+b-1}\sum_{i=1}^{m}p_i\sigma^2 \qquad (37)\\
&=\sum_{a=0}^{K/\tau_{max}-1}\sum_{b=1}^{\tau_{max}}(b-1)\sigma^2=\sum_{a=0}^{K/\tau_{max}-1}\frac{\tau_{max}(\tau_{max}-1)}{2}\sigma^2 \\
&\leq K\frac{(\tau_{max}-1)}{2}\sigma^2. \qquad (38)
\end{aligned}
$$

Remember FedLAMA synchronizes the whole parameters at least once after every τ_{max} iterations. Thus, (35) holds because $\prod_{l=s}^{a\tau_{max}+b-1}\mathbf{W}_l$ is $\mathbf{J}$ when $s\leq a\tau_{max}$, and thus $\mathbf{J}-\prod_{l=s}^{a\tau_{max}+b-1}\mathbf{W}_l$ becomes 0. (36) holds based on Lemma A.2. (37) holds based on Assumption 3.

Bounding T_4

$$
\begin{aligned}
&\sum_{k=1}^{K-\tau_{max}} \mathbb{E}\left[\left\|\sum_{s=1}^{k-1}(\mathbf{J}-\prod_{l=s}^{k-1}\mathbf{W}_l)\hat{\mathbf{f}}_s\right\|^2\right] \\
&= \sum_{a=0}^{K/\tau_{max}-1}\sum_{b=1}^{\tau_{max}} \mathbb{E}\left[\left\|\sum_{s=1}^{a\tau+b-1}(\mathbf{J}-\prod_{l=s}^{a\tau_{max}+b-1}\mathbf{W}_l)\hat{\mathbf{f}}_s\right\|^2\right] \\
&= \sum_{a=0}^{K/\tau_{max}-1}\sum_{b=1}^{\tau_{max}} \mathbb{E}\left[\left\|\sum_{s=a\tau_{max}+1}^{a\tau_{max}+b-1}(\mathbf{J}-\prod_{l=s}^{a\tau_{max}+b-1}\mathbf{P}_l)\hat{\mathbf{f}}_s\right\|^2\right] &(39)\\
&\leq \sum_{a=0}^{K/\tau_{max}-1}\sum_{b=1}^{\tau_{max}}\left((b-1)\sum_{s=a\tau_{max}+1}^{a\tau_{max}+b-1}\mathbb{E}\left[\left\|(\mathbf{J}-\prod_{l=s}^{a\tau_{max}+b-1}\mathbf{P}_l)\hat{\mathbf{f}}_s\right\|^2\right]\right) &(40)\\
&\leq \sum_{a=0}^{K/\tau_{max}-1}\sum_{b=1}^{\tau_{max}}\left((b-1)\sum_{s=a\tau_{max}+1}^{a\tau_{max}+b-1}\mathbb{E}\left[\left\|\hat{\mathbf{f}}_s\right\|^2\left\|(\mathbf{J}-\prod_{l=s}^{a\tau_{max}+b-1}\mathbf{P}_l)\right\|_{op}^2\right]\right) &(41)\\
&\leq \sum_{a=0}^{K/\tau_{max}-1}\sum_{b=1}^{\tau_{max}}\left((b-1)\sum_{s=a\tau_{max}+1}^{a\tau_{max}+b-1}\mathbb{E}\left[\left\|\hat{\mathbf{f}}_s\right\|^2\right]\right) &(42)\\
&\leq \frac{\tau_{max}(\tau_{max}-1)}{2}\sum_{a=0}^{K/\tau_{max}-1}\left(\sum_{s=a\tau_{max}+1}^{a\tau_{max}+\tau_{max}-1}\mathbb{E}\left[\left\|\hat{\mathbf{f}}_s\right\|^2\right]\right)\\
&\leq \frac{\tau_{max}(\tau_{max}-1)}{2}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\hat{\mathbf{f}}_k\right\|^2\right]\\
&= \frac{\tau_{max}(\tau_{max}-1)}{2}\sum_{k=1}^{K}\sum_{i=1}^{m}p_i\,\mathbb{E}\left[\left\|\nabla F_i(\mathbf{x}_k^i)\right\|^2\right], &(43)
\end{aligned}
$$

where (39) holds because $\mathbf{J}-\prod_{l=s}^{a\tau_{max}+b-1}\mathbf{P}_l$ becomes 0 when $s \leq a\tau$. (40) holds based on the convexity of $\ell 2$ norm and Jensen's inequality. (41) holds based on Lemma A.1. (42) holds based on Lemma A.2.

Final Result

By plugging in (38) and (43) into (32), we have

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{m}p_i\,\mathbb{E}\left[\left\|\mathbf{u}_k-\mathbf{x}_k^i\right\|^2\right]\\
&\quad\leq \frac{2\eta^2}{K}\left(K\frac{(\tau_{max}-1)}{2}\sigma^2+\frac{\tau_{max}(\tau_{max}-1)}{2}\left(\sum_{k=1}^{K}\sum_{i=1}^{m}p_i\,\mathbb{E}\left[\left\|\nabla F_i(\mathbf{x}_k^i)\right\|^2\right]\right)\right)\\
&\quad= \eta^2(\tau_{max}-1)\sigma^2+\frac{\eta^2\tau_{max}(\tau_{max}-1)}{K}\left(\sum_{k=1}^{K}\sum_{i=1}^{m}p_i\,\mathbb{E}\left[\left\|\nabla F_i(\mathbf{x}_k^i)\right\|^2\right]\right) &(44)
\end{aligned}
$$

The local gradient term on the right-hand side in (44) can be rewritten using the following inequality.

$$
\begin{aligned}
\mathbb{E}\left[\left\|\nabla F_i(\mathbf{x}_k^i)\right\|^2\right] &= \mathbb{E}\left[\left\|\nabla F_i(\mathbf{x}_k^i)-\nabla F_i(\mathbf{u}_k)+\nabla F_i(\mathbf{u}_k)\right\|^2\right]\\
&\leq 2\,\mathbb{E}\left[\left\|\nabla F_i(\mathbf{x}_k^i)-\nabla F_i(\mathbf{u}_k)\right\|^2\right]+2\,\mathbb{E}\left[\left\|\nabla F_i(\mathbf{u}_k)\right\|^2\right] &(45)\\
&\leq 2L^2\,\mathbb{E}\left[\left\|\mathbf{u}_k-\mathbf{x}_k^i\right\|^2\right]+2\,\mathbb{E}\left[\left\|\nabla F_i(\mathbf{u}_k)\right\|^2\right], &(46)
\end{aligned}
$$

where (45) holds based on the convexity of $\ell 2$ norm and Jensen's inequality.

Plugging in (46) into (44), we have

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{m} p_i \,\mathbb{E}\left[\left\|\mathbf{u}_k - \mathbf{x}_k^i\right\|^2\right] \\
&\quad\leq \eta^2(\tau_{max}-1)\sigma^2 + \frac{2\eta^2\tau_{max}(\tau_{max}-1)L^2}{K}\sum_{k=1}^{K}\sum_{i=1}^{m} p_i \,\mathbb{E}\left[\left\|\mathbf{u}_k - \mathbf{x}_k^i\right\|^2\right] \\
&\qquad + \frac{2\eta^2\tau_{max}(\tau_{max}-1)}{K}\sum_{k=1}^{K}\sum_{i=1}^{m} p_i \,\mathbb{E}\left[\left\|\nabla F_i(\mathbf{u}_k)\right\|^2\right] \qquad (47)
\end{aligned}$$

After a minor rearranging, we have

$$\begin{aligned}
\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{m} p_i \,\mathbb{E}\left[\left\|\mathbf{u}_k - \mathbf{x}_k^i\right\|^2\right] &\leq \frac{\eta^2(\tau_{max}-1)\sigma^2}{1-2\eta^2\tau_{max}(\tau_{max}-1)L^2} \\
&\quad + \frac{2\eta^2\tau_{max}(\tau_{max}-1)}{K(1-2\eta^2\tau_{max}(\tau_{max}-1)L^2)}\sum_{k=1}^{K}\sum_{i=1}^{m} p_i \,\mathbb{E}\left[\left\|\nabla F_i(\mathbf{u}_k)\right\|^2\right] \qquad (48)
\end{aligned}$$

Let us define $A = 2\eta^2\tau_{max}(\tau_{max}-1)L^2$. Then (48) is simplified as follows.

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{m} p_i \,\mathbb{E}\left[\left\|\mathbf{u}_k - \mathbf{x}_k^i\right\|^2\right] \\
&\quad\leq \frac{\eta^2(\tau_{max}-1)\sigma^2}{1-A} + \frac{A}{KL^2(1-A)}\sum_{k=1}^{K}\sum_{i=1}^{m} p_i \,\mathbb{E}\left[\left\|\nabla F_i(\mathbf{u}_k)\right\|^2\right]
\end{aligned}$$

Based on Assumption 4, we have

$$\begin{aligned}
&\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{m} p_i \,\mathbb{E}\left[\left\|\mathbf{u}_k - \mathbf{x}_k^i\right\|^2\right] \\
&\quad\leq \frac{\eta^2(\tau_{max}-1)\sigma^2}{1-A} + \frac{A\beta^2}{KL^2(1-A)}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\sum_{i=1}^{m} p_i \nabla F_i(\mathbf{u}_k)\right\|^2\right] + \frac{A\kappa^2}{L^2(1-A)} \qquad (49) \\
&\quad= \frac{\eta^2(\tau_{max}-1)\sigma^2}{1-A} + \frac{A\beta^2}{KL^2(1-A)}\sum_{k=1}^{K}\mathbb{E}\left[\left\|\nabla F(\mathbf{u}_k)\right\|^2\right] + \frac{A\kappa^2}{L^2(1-A)}, \qquad (50)
\end{aligned}$$

where (50) holds based on the definition of the objective function (10).

Note that (49) is true only when $1 - A > 0$. Thus, after a minor rearrangement, we have a learning rate constraint as follows.

$$\eta < \frac{1}{2(\tau_{max}-1)L} \qquad (51)$$

Here, we complete the proof. □

A.1.3 Proof of Other Lemmas

Lemma A.1. *Consider a real matrix $\mathbf{A} \in \mathbb{R}^{md_j \times md_j}$ and a real vector $\mathbf{b} \in \mathbb{R}^{md_j}$. If $\mathbf{b} \neq \mathbf{0}_{md_j}$, we have*

$$\|\mathbf{A}\mathbf{b}\| \leq \|\mathbf{A}\|_{op}\|\mathbf{b}\| \qquad (52)$$

Proof.

$$\begin{aligned}\|\mathbf{Ab}\|^2 &= \frac{\|\mathbf{Ab}\|^2}{\|\mathbf{b}\|^2}\|\mathbf{b}\|^2 \\ &\leq \|\mathbf{A}\|_{op}^2\|\mathbf{b}\|^2 \end{aligned} \tag{53}$$

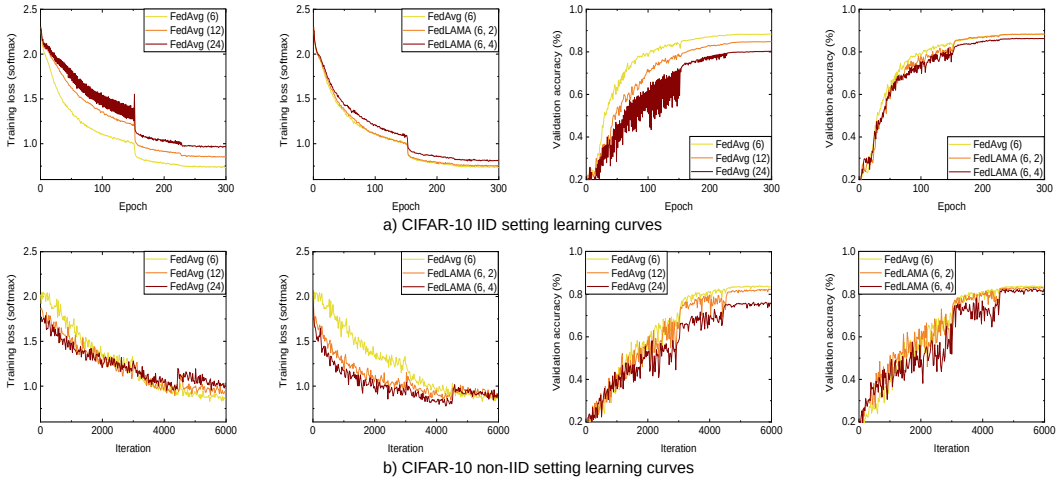
where (53) holds based on the definition of operator norm. □

Lemma A.2. *Suppose an $md \times md$ averaging matrix $\mathbf{P}$ and the full-averaging matrix $\mathbf{J}$, then*

$$\|\mathbf{J} - \mathbf{P}\|_{op}^2 = 1. \tag{54}$$

regardless of which layers are chosen as the LCL.

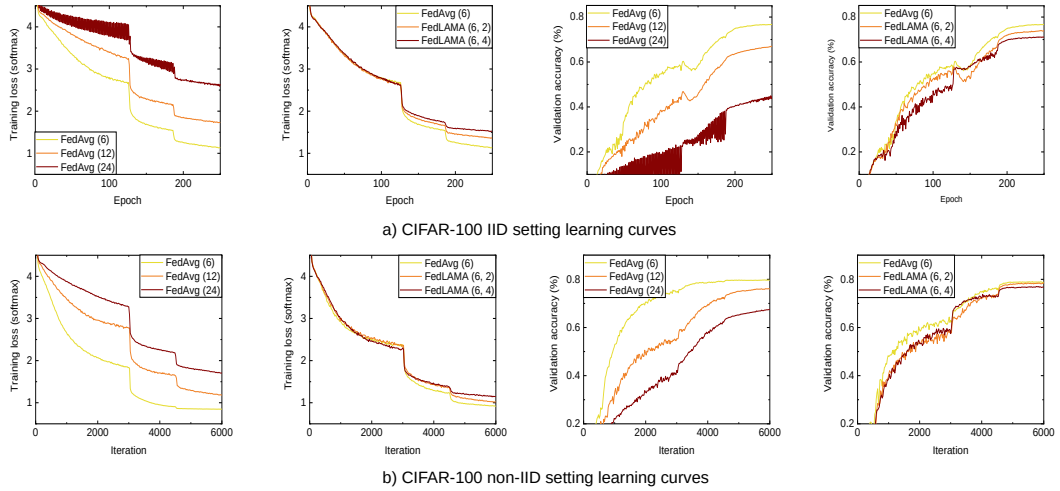
Proof. First, by the definition of averaging matrix $\mathbf{P}$, all the columns that do not correspond to the LCL are zeroed out in $\mathbf{J} - \mathbf{P}$. Then, based on the averaging matrix property 1 and 2, the remaining columns in $\mathbf{P}$ has 1 at all different rows. By the definition of $\mathbf{J}$, all the non-zero elements in i^{th} column are the same $p_i, i \in \{1, \cdots, m\}$. Consequently, the remaining columns in $\mathbf{J} - \mathbf{P}$ are always orthogonal regardless of which layers are chosen as the LCL, and thus the eigenvalues of $\mathbf{J} - \mathbf{P}$ are either 1 or -1. Finally, by the definition of the matrix operator norm, $\|\mathbf{J} - \mathbf{P}\|_{op}^2 = max\{|\lambda(\mathbf{J} - \mathbf{P})|\} = 1$, where $\lambda(\cdot)$ indicates the eigenvalues of the input matrix. □



a) CIFAR-10 IID setting learning curves

b) CIFAR-10 non-IID setting learning curves

Figure 4: The learning curves of CIFAR-10 (ResNet20) training (128 clients). a): The curves for IID data distribution. b): The curves for non-IID data distribution ($\alpha = 0.1$). FedAvg (x) indicates FedAvg with the interval of x. FedLAMA (x, y) indicates FedLAMA with the base interval of x and the interval increase factor of y. As the aggregation interval increases, FedAvg rapidly loses the convergence speed, and it results in achieving a lower validation accuracy within the fixed iteration budget. In contrast, FedLAMA effectively increases the aggregation interval while maintaining the convergence speed.



a) CIFAR-100 IID setting learning curves

b) CIFAR-100 non-IID setting learning curves

Figure 5: The learning curves of CIFAR-100 (WideResNet28-10) training (128 clients). a): The curves for IID data distribution. b): The curves for non-IID data distribution ($\alpha = 0.1$). FedAvg (x) indicates FedAvg with the interval of x. FedLAMA (x, y) indicates FedLAMA with the base interval of x and the interval increase factor of y. While FedAvg significantly loses the convergence speed as the aggregation interval increases, FedLAMA has a marginl impact on it which results in a higher validation accuracy.

A.2 Additional Experimental Results

In this section, we provide extra experimental results with extensive hyper-parameter settings. We commonly use 128 clients and a local batch size of 32 in all the experiments. The gradual learning rate warmup (Goyal et al. (2017)) is also applied to the first 10 epochs in all the experiments. Overall, the learning curve charts and the validation accuracy tables deliver the key insight that FedLAMA achieves a comparable convergence speed to the periodic full aggregation with the base interval

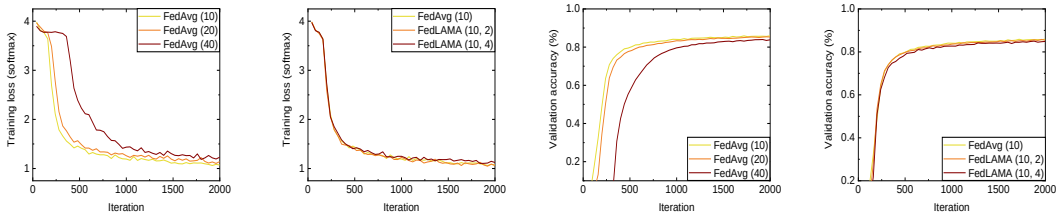


Figure 6: The learning curves of FEMNIST (CNN) training. FedAvg (x) indicates FedAvg with the interval of x. FedLAMA (x, y) indicates FedLAMA with the base interval of x and the interval increase factor of y. FedLAMA curves are not strongly affected by the increased aggregation interval while FedAvg significantly loses the convergence speed as well as the validation accuracy.

(τ') while having the communication cost that is similar to the periodic full aggregation with the increased interval ($\phi\tau'$).

Artificial Data Heterogeneity – For CIFAR-10 and CIFAR-100, we artificially generate the heterogeneous data distribution using Dirichlet's distribution. The concentration coefficient α is set to 0.1, 0.5, and 1.0 to evaluate the performance of FedLAMA across a variety of degree of data heterogeneity. Note that the small concentration coefficient represents the highly heterogeneous numbers of local samples across clients as well as the balance of the samples across the labels. We used the data distribution source code provided by FedML (He et al. (2020)).

CIFAR-10 – Figure 4 shows the full learning curves for IID and non-IID CIFAR-10 datasets. The hyper-parameter settings correspond to Table 4 and 1. First, as the aggregation interval increases from 6 to 24, FedAvg suffers from the slower convergence, and it results in achieving a lower validation accuracy, regardless of the data distribution. In contrast, FedLAMA learning curves are marginally affected by the increased aggregation interval. Table 6 and 7 show the CIFAR-10 classification performance of FedLAMA across different ϕ settings. As expected, the accuracy is reduced as ϕ increases. The IID and non-IID data settings show the common trend. Depending on the system network bandwidth, ϕ can be tuned to be an appropriate value. When $\phi = 2$, the accuracy is almost the same as or even slightly higher than FedAvg accuracy. If the network bandwidth is limited, one can increase ϕ and slightly increase the epoch budget to achieve a good accuracy. Table 8 shows the CIFAR-10 accuracy across different τ' settings. We see that the accuracy is significantly dropped as τ' increases.

CIFAR-100 – Figure 5 shows the learning curves for IID and non-IID CIFAR-100 datasets. Likely to CIFAR-10 results, FedAvg learning curves are strongly affected as the aggregation interval increases from 6 to 24 while FedLAMA learning curves are not strongly affected. Table 9 and 10 show the CIFAR-100 classification performance of FedLAMA across different ϕ settings. FedLAMA achieves a comparable accuracy to FedAvg with a short aggregation interval, even when the degree of data heterogeneity is extreamly high (25% device sampling and Direchlet's coefficient of 0.1). Table 11 shows the FedAvg accuracy with different τ' settings. Under the strongly heterogeneous data distributions, FedAvg with a large aggregation interval ($\tau \geq 12$) do not achieve a reasonable accuracy.

FEMNIST – Figure 6 shows the learning curves of CNN training. Likely to the previous two datasets, the periodic full aggregation suffers from the slower convergence as the aggregation interval increases. FedLAMA learning curves are not much affected by the increased aggregation interval, and it results in achieving a higher validation accuracy after the same number of iterations. Table 12 shows the FEMNIST classification performance of FedLAMA across different ϕ settings. FedLAMA achieves a similar accuracy to the baseline (FedAvg with $\tau' = 10$) even when using a large interval increase factor $\phi \geq 4$. These results demonstrate the effectiveness of the proposed layer-wise adaptive model aggregation method on the problems with heterogeneous data distributions.

Table 6: (IID data) CIFAR-10 classification results of FedLAMA with different ϕ settings.

# of clients	Local batch size	LR	Averaging interval: τ'	Interval increase factor: ϕ	Validation acc.
128	32	0.8	6	1 (FedAvg)	$88.37 \pm 0.1\%$
		0.5		2	$88.41 \pm 0.04\%$
				4	$86.33 \pm 0.2\%$
				8	$85.08 \pm 0.04\%$

Table 7: (Non-IID data) CIFAR-10 classification results of FedLAMA with different ϕ settings.

# of clients	Local batch size	LR	τ'	Active ratio	Dirichlet coeff.	ϕ	Validation acc.
128	32	0.8	6	100%	1	1 (FedAvg)	$90.79 \pm 0.1\%$
						2	$89.01 \pm 0.04\%$
						4	$87.84 \pm 0.01\%$
				100%	0.5	1 (FedAvg)	$90.53 \pm 0.18\%$
						2	$89.21 \pm 0.2\%$
						4	$86.68 \pm 0.12\%$
				100%	0.1	1 (FedAvg)	$89.52 \pm 0.11\%$
						2	$89.00 \pm 0.1\%$
						4	$84.82 \pm 0.08\%$
				50%	1	1 (FedAvg)	$90.34 \pm 0.12\%$
						2	$89.56 \pm 0.13\%$
						4	$87.48 \pm 0.21\%$
				50%	0.5	1 (FedAvg)	$89.86 \pm 0.13\%$
						2	$88.44 \pm 0.15\%$
						4	$87.29 \pm 0.18\%$
				50%	0.1	1 (FedAvg)	$87.83 \pm 0.2\%$
						2	$87.40 \pm 0.17\%$
						4	$85.92 \pm 0.21\%$
		0.6		25%	1	1 (FedAvg)	$88.97 \pm 0.03\%$
						2	$87.89 \pm 0.2\%$
						4	$86.61 \pm 0.1\%$
				25%	0.5	1 (FedAvg)	$87.59 \pm 0.05\%$
						2	$87.12 \pm 0.08\%$
						4	$86.57 \pm 0.02\%$
		0.3		25%	0.1	1 (FedAvg)	$84.02 \pm 0.04\%$
						2	$83.55 \pm 0.02\%$
						4	$83.06 \pm 0.03\%$

Table 8: (Non-IID data) CIFAR-10 classification results of FedAvg with different τ' settings.

# of clients	Local batch size	LR	τ'	Active ratio	Dirichlet coeff.	ϕ	Validation acc.
128	32	0.8	6	100%	0.1	1 (FedAvg)	$89.52 \pm 0.11\%$
			12			1 (FedAvg)	$87.29 \pm 0.05\%$
			24			1 (FedAvg)	$84.82 \pm 0.1\%$
128	32	0.3	6	25%	0.1	1 (FedAvg)	$84.02 \pm 0.1\%$
			12			1 (FedAvg)	$82.48 \pm 0.2\%$
			24			1 (FedAvg)	$76.72 \pm 0.1\%$

Table 9: (IID data) CIFAR-100 classification results of FedLAMA with different ϕ settings.

# of clients	Local batch size	LR	Averaging interval: τ'	Interval increase factor: ϕ	Validation acc.
128	32	0.6	6	1 (FedAvg)	$76.50 \pm 0.02\%$
				2	$75.99 \pm 0.03\%$
				4	$76.17 \pm 0.2\%$
				8	$76.15 \pm 0.2\%$

Table 10: (Non-IID data) CIFAR-100 classification results of FedLAMA with different ϕ settings.

# of clients	Local batch size	LR	τ'	Active ratio	Dirichlet coeff.	ϕ	Validation acc.
128	32	0.4	6	100%	1	1 (FedAvg)	$80.34 \pm 0.01\%$
						2	$78.92 \pm 0.01\%$
						4	$77.16 \pm 0.05\%$
				100%	0.5	1 (FedAvg)	$80.19 \pm 0.02\%$
						2	$78.88 \pm 0.1\%$
						4	$78.03 \pm 0.08\%$
		0.2		100%	0.1	1 (FedAvg)	$79.78 \pm 0.02\%$
						2	$79.07 \pm 0.02\%$
						4	$79.32 \pm 0.01\%$
		0.4		50%	1	1 (FedAvg)	$79.94 \pm 0.1\%$
						2	$78.98 \pm 0.01\%$
						4	$77.50 \pm 0.02\%$
				50%	0.5	1 (FedAvg)	$79.95 \pm 0.05\%$
						2	$78.37 \pm 0.05\%$
						4	$76.93 \pm 0.1\%$
		0.2		50%	0.1	1 (FedAvg)	$79.62 \pm 0.06\%$
						2	$78.76 \pm 0.02\%$
						4	$77.44 \pm 0.02\%$
		0.4		25%	1	1 (FedAvg)	$78.78 \pm 0.02\%$
						2	$78.10 \pm 0.02\%$
		0.2				4	$76.84 \pm 0.03\%$
		0.4		25%	0.5	1 (FedAvg)	$78.81 \pm 0.01\%$
						2	$77.86 \pm 0.04\%$
						4	$77.01 \pm 0.1\%$
		0.2		25%	0.1	1 (FedAvg)	$79.06 \pm 0.03\%$
						2	$78.63 \pm 0.02\%$
						4	$77.17 \pm 0.01\%$

Table 11: (Non-IID data) CIFAR-100 classification results of FedAvg with different τ' settings.

# of clients	Local batch size	LR	τ'	Active ratio	Dirichlet coeff.	ϕ	Validation acc.
128	32	0.4	6	100%	0.1	1 (FedAvg)	$79.78 \pm 0.02\%$
			12			1 (FedAvg)	$77.71 \pm 0.1\%$
			24			1 (FedAvg)	$69.63 \pm 0.1\%$
128	32	0.4	6	25%	0.1	1 (FedAvg)	$79.06 \pm 0.03\%$
			12			1 (FedAvg)	$76.16 \pm 0.05\%$
			24			1 (FedAvg)	$67.43 \pm 0.1\%$

Table 12: FEMNIST classification results of FedLAMA with different ϕ settings.

# of clients	Local batch size	LR	Averaging interval: τ'	Active ratio	Interval increase factor: ϕ	Validation acc.
128	32	0.04	12	100%	1 (FedAvg)	$85.74 \pm 0.21\%$
					2	$85.40 \pm 0.13\%$
					4	$84.67 \pm 0.1\%$
					8	$84.15 \pm 0.18\%$
				50%	1 (FedAvg)	$86.59 \pm 0.2\%$
					2	$86.07 \pm 0.1\%$
					4	$85.77 \pm 0.15\%$
					8	$85.31 \pm 0.03\%$
				25%	1 (FedAvg)	$86.04 \pm 0.2\%$
					2	$86.01 \pm 0.1\%$
					4	$85.62 \pm 0.08\%$
					8	$85.23 \pm 0.1\%$