Appendix
FedAPA: Server-side Gradient-Based Adaptive Personalized Aggregation for Federated Learning on Heterogeneous Data

A CONVERGENCE ANALYSIS

Additional Notation

Here, additional notations are introduced to better represent the process of local model update. Let $f_i(\theta_i)$ denote the embedding function for the i-th client, which may vary across different clients. The decision function for all clients is $h_i(z_i)$. Thus, the labeling function can be written as $F_i(\theta_i, \phi_i) = h_i(\phi_i) \circ f_i(\theta_i)$, and sometimes we use ω_i to represent $[\theta_i; \phi_i]$ for brevity. Let A_i denote the aggregated weight vector of user i, and let $a_{i,j}$ denote the j-th learnable weight parameter of user i's aggregation weight, where $A_i = [a_{i,1}, a_{i,2}, \ldots, a_{i,M}]^T$, for $i, j = 1, 2, \ldots, M$. Therefore, the local loss function of client i can be written as:

$$\mathcal{L}(\omega_i; x, y) = \mathcal{L}(h(\phi_i; f(\theta_i; x)), y) \tag{1}$$

we use t to represent the communication round and $e \in \{\frac{1}{2}, 1, 2, \ldots, E\}$ to represent the local iterations. There are E local iterations in total, so $tE + e$ refers to the e-th local iteration in the communication round $t + 1$. Moreover, tE represents the time step before the aggregation of client parameters, and $tE + 1/2$ represents the time step between parameters aggregation and the first iteration of the current round.

Key Lemmas

Lemma 1. Let Assumption 1 and Assumption 2 in the main text hold. From the beginning of communication round $t + 1$ to the last local update step, the loss function of an arbitrary client can be bounded as:

$$\begin{aligned}\mathbb{E}[\mathcal{L}_{(t+1)E}] \le& \mathbb{E}[\mathcal{L}_{tE+\frac{1}{2}}] + \frac{L_1 E\alpha^2}{2}\sigma^2 \\ &- (\alpha - \frac{L_1\alpha^2}{2})\sum_{e=\frac{1}{2}}^{E-1} ||\nabla\mathcal{L}_{tE+e}||_2^2\end{aligned} \tag{2}$$

Proof. Due to the fact that this lemma applies to an arbitrary client, the client notation i is omitted. Let $\omega^{(t+1)} = \omega^{(t)} - \alpha gr^{(t)}$, then

$$\begin{aligned}\mathcal{L}_{tE+1} &\overset{(a)}{\le} \mathcal{L}_{tE+\frac{1}{2}} + \langle\nabla\mathcal{L}_{tE+\frac{1}{2}}, (\omega^{(tE+1)} - \omega^{(tE+\frac{1}{2})})\rangle \\ &\quad + \frac{L_1}{2}||\omega^{(tE+1)} - \omega^{(tE+\frac{1}{2})}||_2^2 \\ &= \mathcal{L}_{tE+\frac{1}{2}} - \alpha\langle\nabla\mathcal{L}_{tE+\frac{1}{2}}, gr^{(tE+\frac{1}{2})}\rangle \\ &\quad + \frac{L_1}{2}||\alpha gr^{(tE+\frac{1}{2})}||_2^2,\end{aligned} \tag{3}$$

where (a) follows from the quadratic L_1-$Lipschitz$ smooth bound in Assumption 1 in the main text. Taking expectation of both sides of the above equation on the random variable $\xi^{(tE+\frac{1}{2})}$, we have

$$\begin{aligned}\mathbb{E}[\mathcal{L}_{tE+1}] &\le \mathbb{E}[\mathcal{L}_{tE+\frac{1}{2}}] - \alpha\mathbb{E}[\langle\nabla\mathcal{L}_{(tE+\frac{1}{2})}, gr^{(tE+\frac{1}{2})}\rangle] \\ &\quad + \frac{L_1\alpha^2}{2}\mathbb{E}[||gr^{(tE+\frac{1}{2})}||_2^2] \\ &\overset{(b)}{=} \mathbb{E}[\mathcal{L}_{tE+\frac{1}{2}}] - \alpha||\nabla\mathcal{L}_{tE+\frac{1}{2}}||_2^2 \\ &\quad + \frac{L_1\alpha^2}{2}\mathbb{E}[||gr^{(tE+\frac{1}{2})}||_2^2] \\ &\overset{(c)}{\le} \mathbb{E}[\mathcal{L}_{tE+\frac{1}{2}}] - \alpha||\nabla\mathcal{L}_{tE+\frac{1}{2}}||_2^2 \\ &\quad + \frac{L_1\alpha^2}{2}(||\nabla\mathcal{L}_{tE+\frac{1}{2}}||_2^2 + Var(gr^{(tE+\frac{1}{2})})) \\ &= \mathbb{E}[\mathcal{L}_{tE+\frac{1}{2}}] - (\alpha - \frac{L_1\alpha^2}{2})||\nabla\mathcal{L}_{tE+\frac{1}{2}}||_2^2 \\ &\quad + \frac{L_1\alpha^2}{2}Var(gr^{(tE+\frac{1}{2})}) \\ &\overset{(d)}{\le} \mathbb{E}[\mathcal{L}_{tE+\frac{1}{2}}] - (\alpha - \frac{L_1\alpha^2}{2})||\nabla\mathcal{L}_{tE+\frac{1}{2}}||_2^2 \\ &\quad + \frac{L_1\alpha^2}{2}\sigma^2\end{aligned} \tag{4}$$

where (b) and (d) follow from Assumption 2 in the main text, and (c) follows from $Var(x) = \mathbb{E}[x^2] - (\mathbb{E}[x])^2$. Then, by telescoping over E steps, we have

$$\begin{aligned}\mathbb{E}[\mathcal{L}_{(t+1)E}] \le& \mathbb{E}[\mathcal{L}_{tE+\frac{1}{2}}] + \frac{L_1 E\alpha^2}{2}\sigma^2 \\ &- (\alpha - \frac{L_1\alpha^2}{2})\sum_{e=\frac{1}{2}}^{E-1} ||\nabla\mathcal{L}_{tE+e}||_2^2\end{aligned} \tag{5}$$

Lemma 2. For any $t = 1, 2, \ldots,$we have

$$||(A_i^{(t+1)} - e_i)|| \le ||(A_i^{(0)} - e_i)|| + 2\eta Max\sum_{k=0}^{t} ||(\Theta^{(k)})^T|| \tag{6}$$

Proof. (The norm $|| \cdot ||$ appearing in this lemma refers to the

L_2-norm $||\cdot||_2$.)

$$
\begin{aligned}
&||(A_i^{(t+1)} - e_i)|| \\
&\overset{(a)}{=} ||A_i^{(t)} - \eta(\nabla_{A_i}\bar{\theta}_i^{(t+1)})^T(\theta_i^{(t+1)} - \bar{\theta}_i^{\,(t+1)}) - e_i|| \\
&= ||(A_i^{(t)} - e_i) - \eta(\nabla_{A_i}\bar{\theta}_i^{(t+1)})^T(\theta_i^{(t+1)} - \bar{\theta}_i^{\,(t+1)})|| \\
&\overset{(b)}{\leq} ||(A_i^{(t)} - e_i)|| + ||\eta(\nabla_{A_i}\bar{\theta}_i^{(t+1)})^T(\theta_i^{(t+1)} - \bar{\theta}_i^{\,(t+1)})|| \\
&\overset{(c)}{\leq} ||(A_i^{(t)} - e_i)|| + 2\eta Max||(\nabla_{A_i}\bar{\theta}_i^{(t+1)})^T|| \\
&\overset{(d)}{\leq} ||(A_i^{(t)} - e_i)|| + 2\eta Max||(\Theta^{(t)})^T||
\end{aligned}
\tag{7}
$$

Since inequality (7) holds, for any $t = 1, 2, \ldots,$

$$
||(A_i^{(t+1)} - e_i)|| \leq ||(A_i^{(0)} - e_i)|| + 2\eta Max \sum_{k=0}^{t} ||(\Theta^{(k)})^T|| \tag{8}
$$

where $Max = \max_{\substack{i=1,2,\ldots,M \\ t=1,2,\ldots}} ||\theta_i^{(t)}||$, $\Theta^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_M^{(t)})$, $e_i = (0, \ldots, 1, \ldots, 0)^T$ with 1 at the $i\text{-}th$ position. Here, (a) follows from Equation (5) and Equation (7) in the main text, (b) and (c) follow from $||a-b|| \leq ||a|| + ||b||$, and (d) follows from $\bar{\theta}_i^{(t+1)} = \Theta^{(t)} A_i^{(t)}$.

Lemma 3. Let Assumption 3 in the main text hold. After the parameter aggregation at the server, the loss function of an arbitrary client can be bounded as:

$$
\mathbb{E}[\mathcal{L}_{(t+1)E+\frac{1}{2}}] \leq \mathbb{E}[\mathcal{L}_{(t+1)E}] + 2L_2\eta(t+1)Max^3 \tag{9}
$$

Proof. (The norm $||\cdot||$ appearing in this lemma refers to the L_2-norm $||\cdot||_2$.)

$$
\begin{aligned}
\mathcal{L}_{(t+1)E+\frac{1}{2}} &= \mathcal{L}_{(t+1)E} + \mathcal{L}_{(t+1)E+\frac{1}{2}} - \mathcal{L}_{(t+1)E} \\
&\overset{(a)}{\leq} \mathcal{L}_{(t+1)E} + L_2||\theta_i^{((t+1)E+\frac{1}{2})} - \theta_i^{((t+1)E)}|| \\
&= \mathcal{L}_{(t+1)E} + L_2||\bar{\theta}_i^{(t+2)} - \theta_i^{(t+1)}|| \\
&\overset{(b)}{=} \mathcal{L}_{(t+1)E} + L_2||\sum_{j=1}^{M} a_{i,j}\theta_j^{(t+1)} - \theta_i^{(t+1)}|| \\
&= \mathcal{L}_{(t+1)E} + L_2||\Theta^{(t+1)}A_i^{(t+1)} - \theta_i^{(t+1)}|| \\
&\overset{(c)}{\leq} \mathcal{L}_{(t+1)E} + L_2||\Theta^{(t+1)}(A_i^{(t+1)} - e_i)|| \\
&\overset{(d)}{\leq} \mathcal{L}_{(t+1)E} + L_2 Max||(A_i^{(t+1)} - e_i)|| \\
&\overset{(e)}{\leq} \mathcal{L}_{(t+1)E} + L_2 Max(||(A_i^{(0)} - e_i)|| \\
&\qquad + 2\eta Max \sum_{k=0}^{t} ||(\Theta^{(k)})^T||) \\
&\overset{(f)}{\leq} \mathcal{L}_{(t+1)E} + 2L_2\eta(t+1)Max^3
\end{aligned}
\tag{10}
$$

Taking expectations of random variable ξ on both sides, then

$$
\mathbb{E}[\mathcal{L}_{(t+1)E+\frac{1}{2}}] \leq \mathbb{E}[\mathcal{L}_{(t+1)E}] + 2L_2\eta(t+1)Max^3 \tag{11}
$$

where $Max = \max_{\substack{i=1,2,\ldots,M \\ t=1,2,\ldots}} ||\theta_i^{(t)}||$, $\Theta^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)}, \ldots, \theta_M^{(t)})$, $e_i = (0, \ldots, 1, \ldots, 0)^T$ with 1 at the i-th position. Here, (a) follows from $L2\text{-}Lipschitz$ continuity in Assumption 3 in the main text, (b) follows from Equation (5) in the main text, (c) follows from $\theta_i^{(t+1)} = \Theta^{(t+1)}e_i$, (d) follows from $||\Theta^{(t+1)}|| \leq Max$, (e) follows from lemma 2, and (f) follows from $A_i^{(0)} = e_i$.

Theorem

Theorem 1. Let Assumption 1 to 3 in the main text hold. For an arbitrary client, after every communication round, we have

$$
\begin{aligned}
\mathbb{E}[\mathcal{L}_{(t+1)E+\frac{1}{2}}] \leq \mathbb{E}[\mathcal{L}_{tE+\frac{1}{2}}] - (\alpha - \frac{L_1\alpha^2}{2}) \sum_{e=\frac{1}{2}}^{E-1} ||\nabla \mathcal{L}_{tE+e}||_2^2 \\
+ \frac{L_1 E\alpha^2}{2}\sigma^2 + 2L_2\eta(t+1)Max^3
\end{aligned}
\tag{12}
$$

where $Max = \max_{\substack{i=1,2,\ldots,M \\ t=1,2,\ldots}} ||\theta_i^{(t)}||$.

Proof. Combining Lemma 1 and lemma 3, we easily obtain

$$
\begin{aligned}
\mathbb{E}[\mathcal{L}_{(t+1)E+\frac{1}{2}}] \leq \mathbb{E}[\mathcal{L}_{tE+\frac{1}{2}}] - (\alpha - \frac{L_1\alpha^2}{2}) \sum_{e=\frac{1}{2}}^{E-1} ||\nabla \mathcal{L}_{tE+e}||_2^2 \\
+ \frac{L_1 E\alpha^2}{2}\sigma^2 + 2L_2\eta(t+1)Max^3
\end{aligned}
\tag{13}
$$

Corollary 1. The loss function $\mathcal{L}$ for any arbitrary client exhibits a monotonic decrease with each communication round when

$$
\begin{aligned}
\alpha < 4L_2\eta(t+1)Max^3\Big(\sum_{e=\frac{1}{2}}^{E-1} ||\nabla \mathcal{L}_{tE+e}||_2^2 \\
- ((\sum_{e=\frac{1}{2}}^{E-1} ||\nabla \mathcal{L}_{tE+e}||_2^2)^2 \\
- 4L_1L_2\eta(t+1)Max^3(\sum_{e=\frac{1}{2}}^{E-1} ||\nabla \mathcal{L}_{tE+e}||_2^2 + E\sigma^2))^{\frac{1}{2}}\Big)^{-1}
\end{aligned}
\tag{14}
$$

and

$$
\begin{aligned}
\alpha > 4L_2\eta(t+1)Max^3\Big(\sum_{e=\frac{1}{2}}^{E-1} ||\nabla \mathcal{L}_{tE+e}||_2^2 \\
+ ((\sum_{e=\frac{1}{2}}^{E-1} ||\nabla \mathcal{L}_{tE+e}||_2^2)^2 \\
- 4L_1L_2\eta(t+1)Max^3(\sum_{e=\frac{1}{2}}^{E-1} ||\nabla \mathcal{L}_{tE+e}||_2^2 + E\sigma^2))^{\frac{1}{2}}\Big)^{-1}
\end{aligned}
\tag{15}
$$

and

$$\eta < \frac{(\sum_{e=\frac{1}{2}}^{E-1} ||\nabla \mathcal{L}_{tE+e}||_2^2)^2}{4L_1L_2(t+1)Max^3(\sum_{e=\frac{1}{2}}^{E-1} ||\nabla \mathcal{L}_{tE+e}||_2^2 + E\sigma^2)} \tag{16}$$

Hence, the convergence of $\mathcal{L}$ is proven.