Proof for paper titled Convergence Analysis of Client-Edge-Cloud Hierarchical Split Federated Learning

April 21, 2024

1 Assumptions

Assumption 1. *(L-Smoothness). The loss function $f(x)$ is L-smooth with the Lipschitz constant $0<L<\infty$, i.e., for any x and y, we have*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \tag{1}$$

Assumption 2. *(Variance of SGD). The variances of stochastic gradients for client-side model and server-side model across each layer are bounded*

$$\mathbb{E}_{\xi_n \sim \mathcal{D}_n} \left\| \tilde{\nabla} f_{r,n}(\mathbf{w}) - \nabla f_{r,n}(\mathbf{w}) \right\|^2 \leq \sum_{l=1}^{L_c} \sigma_l^2, \forall \mathbf{w}, \forall n, \tag{2}$$

$$\mathbb{E}_{\xi_n \sim \mathcal{D}_n} \left\| \tilde{\nabla} f_{s,n}(\mathbf{w}) - \nabla f_{s,n}(\mathbf{w}) \right\|^2 \leq \sum_{l=L_c+1}^{L^T} \sigma_l^2, \forall \mathbf{w}, \forall n, \tag{3}$$

where L_c is the number of client-side model layers, and L^T represents the total number of layers for the global model. σ_l^2 denotes the bounded variance for the l-th layer of model.

Assumption 3. *(Second moments). The expectation of the squared l_2-norm of the gradients for client-side models and server-side models across each layer are bounded*

$$\mathbb{E}_{\xi_n \sim \mathcal{D}_n} \|\nabla f_{r,n}(\mathbf{w})\|^2 \leq \sum_{l=1}^{L_c} G_l^2, \forall \mathbf{w}, \forall n, \tag{4}$$

$$\mathbb{E}_{\xi_n \sim \mathcal{D}_n} \|\nabla f_{s,n}(\mathbf{w})\|^2 \leq \sum_{l=L_c+1}^{L^T} G_l^2, \forall \mathbf{w}, \forall n, \tag{5}$$

where G_l^2 denotes the second order moments for the l-th layer of model.

2 Model Update

The evolution of the model parameters $\mathbf{x}_r^{k+1}$, $\mathbf{x}_s^{k+1}$ in the cloud are specified as follows:

$$\mathbf{x}_r^{k+1} = \mathbf{x}_r^k - \eta \sum_{z=1}^{Z} \frac{n^z}{N} \frac{1}{n^z} \sum_{\alpha=0}^{\tau_2-1} \sum_{j=1}^{n^z} \sum_{\beta=0}^{\tau_1-1} \tilde{\nabla} f_{r,j}\left(\mathbf{w}_{r,j}^{k,\alpha,\beta}\right), \tag{6}$$

$$\mathbf{x}_s^{k+1} = \mathbf{x}_s^k - \eta \sum_{z=1}^{Z} \frac{n^z}{N} \frac{1}{n^z} \sum_{\alpha=0}^{\tau_1\tau_2-1} \sum_{j=1}^{n^z} \tilde{\nabla} f_{s,j}\left(\mathbf{w}_{s,j}^{k,\alpha}\right). \tag{7}$$

Moreover, the evolution of the client-side model and server-side model parameters $\mathbf{w}_{r,n}^{k,\alpha,\beta}$, $\mathbf{w}_{s,n}^{k,\gamma}$ for client n are specified as follows:

$$\mathbf{w}_{r,n}^{k,\alpha,\beta} = \mathbf{x}_r^k - \eta \sum_{t_1=0}^{\beta-1} \tilde{\nabla} f_{r,n}\left(\mathbf{w}_{r,n}^{k,\alpha,t_1}\right) - \eta \sum_{t_2=0}^{\alpha-1} \sum_{j=1}^{n^z} \frac{1}{n^z} \sum_{t_1=0}^{\tau_1-1} \tilde{\nabla} f_{r,j}\left(\mathbf{w}_{r,j}^{k,t_2,t_1}\right), \tag{8}$$

$$\mathbf{w}_{s,n}^{k,\gamma} = \mathbf{x}_s^k - \eta \sum_{t=0}^{\gamma-1} \frac{1}{n^z} \sum_{j=1}^{n^z} \tilde{\nabla} f_{s,j}\left(\mathbf{w}_{s,j}^{k,t}\right), \tag{9}$$

where $\gamma = \alpha \times \tau_1 + \beta$.

3 Lemma 1

Lemma 1. *(One Round of Cloud Aggregation). With Assumption 1, the following relationship between* $\mathbf{x}^{k+1}$ *and* $\mathbf{x}^k$*:*

$$\begin{aligned}\mathbb{E}\left[f(\mathbf{x}^{k+1})\right] \leq \mathbb{E}\left[f(\mathbf{x}^{k})\right] + \mathbb{E}\left[\langle \nabla f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k\rangle\right] \\ + \frac{L}{2}\mathbb{E}\left[\|\mathbf{x}_r^{k+1} - \mathbf{x}_r^k\|^2\right] + \frac{L}{2}\mathbb{E}\left[\|\mathbf{x}_s^{k+1} - \mathbf{x}_s^k\|^2\right].\end{aligned} \tag{10}$$

Proof. The proof directly follows from the property of the L-smoothness:

$$\mathbb{E}\left[f(\mathbf{x}^{k+1})\right] \leq \mathbb{E}\left[f(\mathbf{x}^{k})\right] + \mathbb{E}\left[\langle \nabla f(\mathbf{x}^k), \mathbf{x}^{k+1} - \mathbf{x}^k\rangle\right] + \frac{L}{2}\mathbb{E}\left[\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2\right]. \tag{11}$$

Note that,

$$\mathbb{E}\left[\|\mathbf{x}^{k+1} - \mathbf{x}^k\|^2\right] = \mathbb{E}\left[\|\mathbf{x}_r^{k+1} - \mathbf{x}_r^k\|^2\right] + \mathbb{E}\left[\|\mathbf{x}_s^{k+1} - \mathbf{x}_s^k\|^2\right]. \tag{12}$$

Substituting Eq. (12) into Eq. (11), we complete the proof.

4 Lemma 2

Lemma 2. *With Assumptions 1, 2, and 3,* $\mathbb{E}\left[\|\mathbf{x}_r^{k+1} - \mathbf{x}_r^k\|^2\right]$ *is bounded as follows:*

$$\mathbb{E}\|\mathbf{x}_r^{k+1} - \mathbf{x}_r^k\|^2 \leq \frac{\eta^2}{N}\sum_{i=1}^{N}\tau_1\tau_2\sum_{\alpha=0}^{\tau_2-1}\sum_{\beta=0}^{\tau_1-1}\mathbb{E}\|\nabla f_{r,i}\left(\mathbf{w}_{r,i}^{k,\alpha,\beta}\right)\|^2 + \eta^2\tau_1^2\tau_2^2\sum_{l=1}^{L_c}\sigma_l^2. \tag{13}$$

Proof.

$$\begin{aligned}&\mathbb{E}\|\mathbf{x}_r^{k+1} - \mathbf{x}_r^k\|^2 \\ &\overset{(a)}{=}\eta^2\mathbb{E}\|\sum_{i=1}^{N}\frac{1}{N}\sum_{\alpha=0}^{\tau_2-1}\sum_{\beta=0}^{\tau_1-1}\tilde{\nabla} f_{r,i}\left(\mathbf{w}_{r,i}^{k,\alpha,\beta}\right)\|^2 \\ &\overset{(b)}{=}\eta^2\mathbb{E}\|\sum_{i=1}^{N}\frac{1}{N}\sum_{\alpha=0}^{\tau_2-1}\sum_{\beta=0}^{\tau_1-1}\nabla f_{r,i}\left(\mathbf{w}_{r,i}^{k,\alpha,\beta}\right)\|^2 + \eta^2\mathbb{E}\|\sum_{i=1}^{N}\frac{1}{N}\sum_{\alpha=0}^{\tau_2-1}\sum_{\beta=0}^{\tau_1-1}\tilde{\nabla} f_{r,i}\left(\mathbf{w}_{r,i}^{k,\alpha,\beta}\right) - \nabla f_{r,i}\left(\mathbf{w}_{r,i}^{k,\alpha,\beta}\right)\|^2 \\ &\overset{(c)}{\leq}\frac{\eta^2}{N}\sum_{i=1}^{N}\tau_1\tau_2\sum_{\alpha=0}^{\tau_2-1}\sum_{\beta=0}^{\tau_1-1}\mathbb{E}\|\nabla f_{r,i}\left(\mathbf{w}_{r,i}^{k,\alpha,\beta}\right)\|^2 + \frac{\eta^2}{N}\sum_{i=1}^{N}\tau_1\tau_2\sum_{\alpha=0}^{\tau_2-1}\sum_{\beta=0}^{\tau_1-1}\mathbb{E}\|\tilde{\nabla} f_{r,i}\left(\mathbf{w}_{r,i}^{k,\alpha,\beta}\right) - \nabla f_{r,i}\left(\mathbf{w}_{r,i}^{k,\alpha,\beta}\right)\|^2 \\ &\overset{(d)}{\leq}\frac{\eta^2}{N}\sum_{i=1}^{N}\tau_1\tau_2\sum_{\alpha=0}^{\tau_2-1}\sum_{\beta=0}^{\tau_1-1}\mathbb{E}\|\nabla f_{r,i}\left(\mathbf{w}_{r,i}^{k,\alpha,\beta}\right)\|^2 + \eta^2\tau_1^2\tau_2^2\sum_{l=1}^{L_c}\sigma_l^2,\end{aligned} \tag{14}$$

where (a) follows from Eq. (6); (b) follows by applying the Eq. $\mathbb{E}\|x\|^2 = \|\mathbb{E}x\|^2 + \mathrm{Var}(x)^2$; (c) follows from the convexity of $\|\cdot\|_2^2$, and (d) follows from Eq. (2).

5 Lemma 3

Lemma 3. *With Assumptions 1, 2, and 3,* $\mathbb{E}\left[\|\mathbf{x}_s^{k+1} - \mathbf{x}_s^k\|^2\right]$ *is bounded as follows:*

$$\mathbb{E}\|\mathbf{x}_s^{k+1} - \mathbf{x}_s^k\|^2 \leq \frac{\eta^2}{N}\sum_{i=1}^{N}\tau_1\tau_2\sum_{\alpha=0}^{\tau_1\tau_2-1}\mathbb{E}\|\nabla f_{s,i}\left(\mathbf{w}_{s,i}^{k,\alpha}\right)\|^2 + \eta^2\tau_1^2\tau_2^2\sum_{l=L_c+1}^{L^T}\sigma_l^2. \tag{15}$$

The proof process is similar to Lemma 2.

Proof.

$$
\begin{aligned}
&\mathbb{E}\|\mathbf{x}_s^{k+1}-\mathbf{x}_s^{k}\|^2\\
\overset{(a)}{=}&\eta^2\mathbb{E}\|\sum_{i=1}^{N}\frac{1}{N}\sum_{\alpha=0}^{\tau_1\tau_2-1}\tilde{\nabla} f_{s,i}\left(\mathbf{w}_{s,i}^{k,\alpha}\right)\|^2\\
\overset{(b)}{=}&\eta^2\mathbb{E}\|\sum_{i=1}^{N}\frac{1}{N}\sum_{\alpha=0}^{\tau_1\tau_2-1}\nabla f_{s,i}\left(\mathbf{w}_{s,i}^{k,\alpha}\right)\|^2+\eta^2\mathbb{E}\|\sum_{i=1}^{N}\frac{1}{N}\sum_{\alpha=0}^{\tau_1\tau_2-1}\tilde{\nabla} f_{s,i}\left(\mathbf{w}_{s,i}^{k,\alpha}\right)-\nabla f_{s,i}\left(\mathbf{w}_{s,i}^{k,\alpha}\right)\|^2\\
\overset{(c)}{\le}&\frac{\eta^2}{N}\sum_{i=1}^{N}\tau_1\tau_2\sum_{\alpha=0}^{\tau_1\tau_2-1}\mathbb{E}\|\nabla f_{s,i}\left(\mathbf{w}_{s,i}^{k,\alpha}\right)\|^2+\frac{\eta^2}{N}\sum_{i=1}^{N}\tau_1\tau_2\sum_{\alpha=0}^{\tau_1\tau_2-1}\mathbb{E}\|\tilde{\nabla} f_{s,i}\left(\mathbf{w}_{s,i}^{k,\alpha}\right)-\nabla f_{s,i}\left(\mathbf{w}_{s,i}^{k,\alpha}\right)\|^2\\
\overset{(d)}{\le}&\frac{\eta^2}{N}\sum_{i=1}^{N}\tau_1\tau_2\sum_{\alpha=0}^{\tau_1\tau_2-1}\mathbb{E}\|\nabla f_{s,i}\left(\mathbf{w}_{s,i}^{k,\alpha}\right)\|^2+\eta^2\tau_1^2\tau_2^2\sum_{l=L_c+1}^{L^T}\sigma_l^2,
\end{aligned}
\tag{16}
$$

where (a) follows from Eq. (7); (b) follows by applying the Eq. $\mathbb{E}\|x\|^2=\|\mathbb{E}x\|^2+\mathrm{Var}(x)^2$; (c) follows from the convexity of $\|\cdot\|_2^2$, and (d) follows from Eq. (3).

Thus, substituting Eq. (13) and Eq. (15) into Eq. (12) yields

$$
\mathbb{E}\left[\|\mathbf{x}^{k+1}-\mathbf{x}^{k}\|^2\right]=\frac{\eta^2}{N}\sum_{i=1}^{N}\tau_1\tau_2\sum_{\alpha=0}^{\tau_2-1}\sum_{\beta=0}^{\tau_1-1}\mathbb{E}\|\nabla f_i\left(\mathbf{w}_i^{k,\alpha,\beta}\right)\|^2+\eta^2\tau_1^2\tau_2^2\sum_{l=1}^{L^T}\sigma_l^2. \tag{17}
$$

6 Lemma 4

Lemma 4. *With Assumption 1, 2, and 3, $\mathbb{E}\left[\langle\nabla f(\mathbf{x}^k),\mathbf{x}^{k+1}-\mathbf{x}^k\rangle\right]$ is bounded as follows:*

$$
\begin{aligned}
&\mathbb{E}\left[\langle\nabla f(\mathbf{x}^k),\mathbf{x}^{k+1}-\mathbf{x}^k\rangle\right]\\
\le&-\frac{\eta\tau_1\tau_2}{2}\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2-\frac{\eta\tau_1\tau_2}{2}\mathbb{E}\|\frac{1}{N}\sum_{j=1}^{N}\nabla f_j(\mathbf{w}_j^{k,\alpha,\beta})\|^2\\
&+\frac{L^2\eta^3}{2}\tau_1\tau_2C\left(\tau_1,\tau_2\right)\sum_{l=1}^{L^T}\left(G_l^2+\sigma_l^2\right)+\frac{L^2\eta^3}{2}\tau_1\tau_2D\left(\tau_1,\tau_2\right)\sum_{l=1}^{L_c}\left(G_l^2+\sigma_l^2\right),
\end{aligned}
\tag{18}
$$

where $C\left(\tau_1,\tau_2\right)$ and $D\left(\tau_1,\tau_2\right)$ are functions of τ_1 and τ_2,

$$
C\left(\tau_1,\tau_2\right)=\frac{\tau_1^2\left(\tau_2-1\right)\left(2\tau_2-1\right)}{6}+\frac{\left(\tau_1-1\right)\left(2\tau_1-1\right)}{6}+\frac{\tau_1\left(\tau_2-1\right)\left(\tau_1-1\right)}{2}, \tag{19}
$$

$$
D\left(\tau_1,\tau_2\right)=\frac{\tau_1^2\left(\tau_2-1\right)\left(2\tau_2-1\right)}{6}+\frac{\left(\tau_1-1\right)\left(2\tau_1-1\right)}{6}-\frac{\tau_1\left(\tau_2-1\right)\left(\tau_1-1\right)}{2}. \tag{20}
$$

Proof.

By taking the expectation, we obtain:

$$
\begin{aligned}
&\mathbb{E}\left[\langle\nabla f(\mathbf{x}^k),\mathbf{x}^{k+1}-\mathbf{x}^k\rangle\right]\\
=&-\eta\mathbb{E}\left[\langle\nabla f(\mathbf{x}^k),\frac{1}{N}\sum_{j=1}^{N}\sum_{\alpha=0}^{\tau_2-1}\sum_{\beta=0}^{\tau_1-1}\nabla f_j(\mathbf{w}_j^{k,\alpha,\beta})\rangle\right]\\
=&-\eta\sum_{\alpha=0}^{\tau_2-1}\sum_{\beta=0}^{\tau_1-1}\mathbb{E}\left[\langle\nabla f(\mathbf{x}^k),\frac{1}{N}\sum_{j=1}^{N}\nabla f_j(\mathbf{w}_j^{k,\alpha,\beta})\rangle\right]\\
\overset{(a)}{=}&\sum_{\alpha=0}^{\tau_2-1}\sum_{\beta=0}^{\tau_1-1}\left[-\frac{\eta}{2}\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2-\frac{\eta}{2}\mathbb{E}\|\frac{1}{N}\sum_{j=1}^{N}\nabla f_j(\mathbf{w}_j^{k,\alpha,\beta})\|^2+\frac{\eta}{2}\mathbb{E}\|\nabla f(\mathbf{x}^k)-\frac{1}{N}\sum_{j=1}^{N}\nabla f_j(\mathbf{w}_j^{k,\alpha,\beta})\|^2\right],
\end{aligned}
\tag{21}
$$

where (a) follows by using the identity $2\langle\boldsymbol{a},\boldsymbol{b}\rangle=\|\boldsymbol{a}\|^2+\|\boldsymbol{b}\|^2-\|\boldsymbol{a}-\boldsymbol{b}\|^2$.

$$
\begin{aligned}
&\mathbb{E}\|\nabla f(\mathbf{x}_r^k) - \nabla f_{r,j}\left(\mathbf{w}_{r,j}^{k,\alpha,\beta}\right)\|^2 \\
\overset{(a)}{\leq} & L^2\eta^2\mathbb{E}\|\sum_{t_1=0}^{\beta-1}\tilde{\nabla} f_{r,j}\left(\mathbf{w}_{r,j}^{k,\alpha,t_1}\right) + \sum_{t_2=0}^{\alpha-1}\sum_{j=1}^{n^z}\frac{1}{n^z}\sum_{t_1=0}^{\tau_1-1}\tilde{\nabla} f_{r,j}\left(\mathbf{w}_{r,j}^{k,t_2,t_1}\right)\|^2 \\
\overset{(b)}{=} & L^2\eta^2\mathbb{E}\|\sum_{t_1=0}^{\beta-1}\nabla f_{r,j}\left(\mathbf{w}_{r,j}^{k,\alpha,t_1}\right) + \sum_{t_2=0}^{\alpha-1}\sum_{j=1}^{n^z}\frac{1}{n^z}\sum_{t_1=0}^{\tau_1-1}\nabla f_{r,j}\left(\mathbf{w}_{r,j}^{k,t_2,t_1}\right)\|^2 \\
&+ L^2\eta^2\mathbb{E}\|\sum_{t_1=0}^{\beta-1}\left(\tilde{\nabla} f_{r,j}\left(\mathbf{w}_{r,j}^{k,\alpha,t_1}\right) - \nabla f_{r,j}\left(\mathbf{w}_{r,j}^{k,\alpha,t_1}\right)\right) + \sum_{t_2=0}^{\alpha-1}\sum_{j=1}^{n^z}\frac{1}{n^z}\sum_{t_1=0}^{\tau_1-1}\left(\tilde{\nabla} f_{r,j}\left(\mathbf{w}_{r,j}^{k,t_2,t_1}\right) - \nabla f_{r,j}\left(\mathbf{w}_{r,j}^{k,t_2,t_1}\right)\right)\|^2 \\
\overset{(c)}{\leq} & L^2\eta^2\mathbb{E}\|\sum_{t_1=0}^{\beta-1}\nabla f_{r,j}\left(\mathbf{w}_{r,j}^{k,\alpha,t_1}\right) + \sum_{t_2=0}^{\alpha-1}\sum_{j=1}^{n^z}\frac{1}{n^z}\sum_{t_1=0}^{\tau_1-1}\nabla f_{r,j}\left(\mathbf{w}_{r,j}^{k,t_2,t_1}\right)\|^2 \\
&+ 2L^2\eta^2\left(\underbrace{\mathbb{E}\|\sum_{t_1=0}^{\beta-1}\left(\tilde{\nabla} f_{r,j}\left(\mathbf{w}_{r,j}^{k,\alpha,t_1}\right) - \nabla f_{r,j}\left(\mathbf{w}_{r,j}^{k,\alpha,t_1}\right)\right)\|^2}_{A_2} + \underbrace{\mathbb{E}\|\sum_{t_2=0}^{\alpha-1}\sum_{j=1}^{n^z}\frac{1}{n^z}\sum_{t_1=0}^{\tau_1-1}\left(\tilde{\nabla} f_{r,j}\left(\mathbf{w}_{r,j}^{k,t_2,t_1}\right) - \nabla f_{r,j}\left(\mathbf{w}_{r,j}^{k,t_2,t_1}\right)\right)\|^2}_{A_3}\right) \\
\overset{(d)}{\leq} & \underbrace{2L^2\eta^2\beta^2\sum_{l=1}^{L_c}G_l^2 + 2L^2\eta^2\alpha^2\tau_1^2\sum_{l=1}^{L_c}G_l^2}_{A_1} + 2L^2\eta^2(A_2 + A_3),
\end{aligned}
\tag{26}
$$

Now, we will bound the third term on the right hand side (RHS) of Eq. (21).

$$
\begin{aligned}
&\frac{\eta}{2}\mathbb{E}\|\nabla f(\mathbf{x}^k) - \frac{1}{N}\sum_{j=1}^{N}\nabla f_j(\mathbf{w}_j^{k,\alpha,\beta})\|^2 \\
=&\frac{\eta}{2}\mathbb{E}\|\nabla f(\mathbf{x}_r^k) - \frac{1}{N}\sum_{j=1}^{N}\nabla f_{r,j}(\mathbf{w}_{r,j}^{k,\alpha,\beta})\|^2 + \frac{\eta}{2}\mathbb{E}\|\nabla f(\mathbf{x}_s^k) - \frac{1}{N}\sum_{j=1}^{N}\nabla f_{s,j}(\mathbf{w}_{s,j}^{k,\alpha,\beta})\|^2,
\end{aligned}
\tag{22}
$$

where $\frac{\eta}{2}\mathbb{E}\|\nabla f(\mathbf{x}_r^k) - \frac{1}{N}\sum_{j=1}^{N}\nabla f_{r,j}(\mathbf{w}_{r,j}^{k,\alpha,\beta})\|^2$ can be bounded as

$$
\begin{aligned}
&\frac{\eta}{2}\mathbb{E}\|\nabla f(\mathbf{x}_r^k) - \frac{1}{N}\sum_{j=1}^{N}\nabla f_{r,j}(\mathbf{w}_{r,j}^{k,\alpha,\beta})\|^2 \\
=&\frac{\eta}{2}\mathbb{E}\|\frac{1}{N}\sum_{j=1}^{N}\nabla f(\mathbf{x}_r^k) - \frac{1}{N}\sum_{j=1}^{N}\nabla f_{r,j}(\mathbf{w}_{r,j}^{k,\alpha,\beta})\|^2 \\
\leq&\frac{\eta}{2N}\sum_{j=1}^{N}\mathbb{E}\|\nabla f(\mathbf{x}_r^k) - \nabla f_{r,j}(\mathbf{w}_{r,j}^{k,\alpha,\beta})\|^2,
\end{aligned}
\tag{23}
$$

where the bound on the RHS, Eq. (26), shown at the beginning of this page.

Now we continue to bound the two terms A_2 and A_3 in Eq. (26) as

$$
A_2 \leq \beta\sum_{t_1=0}^{\beta-1}\mathbb{E}\|\tilde{\nabla} f_{r,j}\left(\mathbf{w}_{r,j}^{k,\alpha,t_1}\right) - \nabla f_{r,j}\left(\mathbf{w}_{r,j}^{k,\alpha,t_1}\right)\|^2 \leq \beta^2\sum_{l=1}^{L_c}\sigma_l^2. \tag{24}
$$

$$
A_3 \leq \alpha\tau_1\sum_{t_2=0}^{\alpha-1}\sum_{t_1=0}^{\tau_1-1}\sum_{j=1}^{n^z}\frac{1}{n^z}\mathbb{E}\|\tilde{\nabla} f_{r,j}\left(\mathbf{w}_{r,j}^{k,t_2,t_1}\right) - \nabla f_{r,j}\left(\mathbf{w}_{r,j}^{k,t_2,t_1}\right)\|^2 \leq \alpha^2\tau_1^2\sum_{l=1}^{L_c}\sigma_l^2. \tag{25}
$$

Thus, substituting Eq. (26), Eq. (24), and Eq. (25) into Eq. (23) yields

$$
\begin{aligned}
&\frac{\eta}{2}\mathbb{E}\|\nabla f(\mathbf{x}_r^k)-\frac{1}{N}\sum_{j=1}^{N}\nabla f_{r,j}(\mathbf{w}_{r,j}^{k,\alpha,\beta})\|^2\\
\leq&\frac{\eta}{2N}\sum_{j=1}^{N}\left[2L^2\eta^2\beta^2\sum_{l=1}^{L_c}G_l^2+2L^2\eta^2\alpha^2\tau_1^2\sum_{l=1}^{L_c}G_l^2+2L^2\eta^2\sum_{l=1}^{L_c}\sigma_l^2\left(\beta^2+\alpha^2\tau_1^2\right)\right]\\
\leq&L^2\eta^3\sum_{l=1}^{L_c}G_l^2\left(\beta^2+\alpha^2\tau_1^2\right)+L^2\eta^3\sum_{l=1}^{L_c}\sigma_l^2\left(\beta^2+\alpha^2\tau_1^2\right).
\end{aligned}
\tag{27}
$$

Next, $\frac{\eta}{2}\mathbb{E}\|\nabla f(\mathbf{x}_s^k)-\frac{1}{N}\sum_{j=1}^{N}\nabla f_{s,j}(\mathbf{w}_{s,j}^{k,\alpha,\beta})\|^2$ can be bounded as

$$
\begin{aligned}
&\frac{\eta}{2}\mathbb{E}\|\nabla f(\mathbf{x}_s^k)-\frac{1}{N}\sum_{j=1}^{N}\nabla f_{s,j}(\mathbf{w}_{s,j}^{k,\alpha,\beta})\|^2\\
=&\frac{\eta}{2}\mathbb{E}\|\frac{1}{N}\sum_{j=1}^{N}\nabla f(\mathbf{x}_s^k)-\frac{1}{N}\sum_{j=1}^{N}\nabla f_{s,j}(\mathbf{w}_{s,j}^{k,\alpha,\beta})\|^2\\
\leq&\frac{\eta}{2N}\sum_{j=1}^{N}\mathbb{E}\|\nabla f(\mathbf{x}_s^k)-\nabla f_{s,j}(\mathbf{w}_{s,j}^{k,\alpha,\beta})\|^2,
\end{aligned}
\tag{28}
$$

the term on the RHS can be bounded as

$$
\begin{aligned}
&\frac{\eta}{2}\mathbb{E}\|\nabla f(\mathbf{x}_s^k)-\nabla f_{s,j}(\mathbf{w}_{s,j}^{k,\alpha,\beta})\|^2\\
\overset{(a)}{\leq}&\frac{\eta^3}{2}L^2\mathbb{E}\|\sum_{t=0}^{\alpha\tau_1+\beta-1}\sum_{j=1}^{n^z}\frac{1}{n^z}\tilde{\nabla}f_{s,j}\left(\mathbf{w}_{s,j}^{k,t}\right)\|^2\\
\overset{(b)}{=}&\frac{\eta^3}{2}L^2\mathbb{E}\|\sum_{t=0}^{\alpha\tau_1+\beta-1}\sum_{j=1}^{n^z}\frac{1}{n^z}\nabla f_{s,j}\left(\mathbf{w}_{s,j}^{k,t}\right)\|^2+\frac{\eta^3}{2}L^2\mathbb{E}\|\sum_{t=0}^{\alpha\tau_1+\beta-1}\sum_{j=1}^{n^z}\frac{1}{n^z}\left(\tilde{\nabla}f_{s,j}\left(\mathbf{w}_{s,j}^{k,t}\right)-\nabla f_{s,j}\left(\mathbf{w}_{s,j}^{k,t}\right)\right)\|^2\\
\overset{(c)}{\leq}&\frac{\eta^3}{2}L^2\left(\alpha\tau_1+\beta\right)\sum_{t=0}^{\alpha\tau_1+\beta-1}\frac{1}{n^z}\sum_{j=1}^{n^z}\mathbb{E}\|\nabla f_{s,j}\left(\mathbf{w}_{s,j}^{k,t}\right)\|^2+\frac{\eta^3}{2}L^2\left(\alpha\tau_1+\beta\right)^2\sum_{l=L_c+1}^{L^T}\sigma_l^2\\
\overset{(d)}{\leq}&\frac{\eta^3}{2}L^2\left(\alpha\tau_1+\beta\right)^2\sum_{l=L_c+1}^{L^T}G_l^2+\frac{\eta^3}{2}L^2\left(\alpha\tau_1+\beta\right)^2\sum_{l=L_c+1}^{L^T}\sigma_l^2,
\end{aligned}
\tag{29}
$$

where (a) follows from Eq. (9); (b) follows by applying the Eq. $\mathbb{E}\|x\|^2=\|\mathbb{E}x\|^2+\mathrm{Var}(x)^2$; (c) follows from the convexity of $\|\cdot\|_2^2$ and Eq. (3), and (d) follows from Eq. (5). Thus, substituting Eq. (27), Eq. (28), and Eq. (29) into Eq. (22) yields

$$
\begin{aligned}
&\frac{\eta}{2}\mathbb{E}\|\nabla f(\mathbf{x}^k)-\frac{1}{N}\sum_{j=1}^{N}\nabla f_j(\mathbf{w}_j^{k,\alpha,\beta})\|^2\\
=&L^2\eta^3\frac{\left(\alpha\tau_1+\beta\right)^2}{2}\sum_{l=1}^{L^T}\left(G_l^2+\sigma_l^2\right)+L^2\eta^3\frac{\left(\alpha\tau_1-\beta\right)^2}{2}\sum_{l=1}^{L_c}\left(G_l^2+\sigma_l^2\right)
\end{aligned}
\tag{30}
$$

Thus, substituting Eq. (30) into Eq. (21) yields

$$
\begin{aligned}
&\mathbb{E}\left[\langle\nabla f(\mathbf{x}^k), \mathbf{x}^{k+1}-\mathbf{x}^k\rangle\right] \\
\leq& -\frac{\eta\tau_1\tau_2}{2}\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 - \frac{\eta\tau_1\tau_2}{2}\mathbb{E}\|\frac{1}{N}\sum_{j=1}^{N}\nabla f_j(\mathbf{w}_j^{k,\alpha,\beta})\|^2 + \sum_{\alpha=0}^{\tau_2-1}\sum_{\beta=0}^{\tau_1-1}\left(L^2\eta^3\frac{(\alpha\tau_1+\beta)^2}{2}\sum_{l=1}^{L^T}\left(G_l^2+\sigma_l^2\right)\right. \\
&\left.+L^2\eta^3\frac{(\alpha\tau_1-\beta)^2}{2}\sum_{l=1}^{L_c}\left(G_l^2+\sigma_l^2\right)\right) \\
=& -\frac{\eta\tau_1\tau_2}{2}\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 - \frac{\eta\tau_1\tau_2}{2}\mathbb{E}\|\frac{1}{N}\sum_{j=1}^{N}\nabla f_j(\mathbf{w}_j^{k,\alpha,\beta})\|^2 + \frac{L^2\eta^3}{2}\tau_1\tau_2 C\left(\tau_1,\tau_2\right)\sum_{l=1}^{L^T}\left(G_l^2+\sigma_l^2\right) \\
&+\frac{L^2\eta^3}{2}\tau_1\tau_2 D\left(\tau_1,\tau_2\right)\sum_{l=1}^{L_c}\left(G_l^2+\sigma_l^2\right)
\end{aligned}
\tag{31}
$$

where

$$
C\left(\tau_1,\tau_2\right) = \frac{\tau_1^2\left(\tau_2-1\right)\left(2\tau_2-1\right)}{6} + \frac{\left(\tau_1-1\right)\left(2\tau_1-1\right)}{6} + \frac{\tau_1\left(\tau_2-1\right)\left(\tau_1-1\right)}{2} \tag{32}
$$

$$
D\left(\tau_1,\tau_2\right) = \frac{\tau_1^2\left(\tau_2-1\right)\left(2\tau_2-1\right)}{6} + \frac{\left(\tau_1-1\right)\left(2\tau_1-1\right)}{6} - \frac{\tau_1\left(\tau_2-1\right)\left(\tau_1-1\right)}{2} \tag{33}
$$

7 Theorem 1

Theorem 1. *(Convergence of HierSFL for Non-Convex Loss Functions). The mean square gradient of cloud model after K aggregation rounds is bounded:*

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 \\
\leq&\frac{2}{\eta\tau_1\tau_2 K}\left(f(\mathbf{x}^0)-f^*\right) - \sum_{l=1}^{L^T}G_l^2 + L^2\eta^2 C\left(\tau_1,\tau_2\right)\sum_{l=1}^{L^T}\left(G_l^2+\sigma_l^2\right) + L\eta\tau_1\tau_2\sum_{l=1}^{L^T}\left(G_l^2+\sigma_l^2\right) \\
&+L^2\eta^2 D\left(\tau_1,\tau_2\right)\sum_{l=1}^{L_c}\left(G_l^2+\sigma_l^2\right),
\end{aligned}
\tag{34}
$$

where f^* represents the minimum value.

Proof.
By combining Lemmas 1 to 4, we now have the following:

$$
\begin{aligned}
&\mathbb{E}\left[f(\mathbf{x}^{k+1})\right] \\
\leq&\mathbb{E}\left[f(\mathbf{x}^k)\right] - \frac{\eta\tau_1\tau_2}{2}\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 - \frac{\eta\tau_1\tau_2}{2}\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}\|\nabla f_j(\mathbf{w}_j^{k,\alpha,\beta})\|^2 + \frac{L\eta^2}{2N}\sum_{j=1}^{N}\tau_1\tau_2\sum_{\alpha=0}^{\tau_2-1}\sum_{\beta=0}^{\tau_1-1}\mathbb{E}\|\nabla f_j\left(\mathbf{w}_j^{k,\alpha,\beta}\right)\|^2 \\
&+\frac{L^2\eta^3}{2}\tau_1\tau_2 C\left(\tau_1,\tau_2\right)\sum_{l=1}^{L^T}\left(G_l^2+\sigma_l^2\right) + \frac{L^2\eta^3}{2}\tau_1\tau_2 D\left(\tau_1,\tau_2\right)\sum_{l=1}^{L_c}\left(G_l^2+\sigma_l^2\right) + \frac{L}{2}\eta^2\tau_1^2\tau_2^2\sum_{l=1}^{L^T}\sigma_l^2 \\
\leq&\mathbb{E}\left[f(\mathbf{x}^k)\right] - \frac{\eta\tau_1\tau_2}{2}\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 - \frac{\eta\tau_1\tau_2}{2}\sum_{l=1}^{L^T}G_l^2 + \frac{L^2\eta^3}{2}\tau_1\tau_2 C\left(\tau_1,\tau_2\right)\sum_{l=1}^{L^T}\left(G_l^2+\sigma_l^2\right) + \frac{L\eta^2}{2}\tau_1^2\tau_2^2\sum_{l=1}^{L^T}G_l^2 \\
&+\frac{L^2\eta^3}{2}\tau_1\tau_2 D\left(\tau_1,\tau_2\right)\sum_{l=1}^{L_c}\left(G_l^2+\sigma_l^2\right) + \frac{L}{2}\eta^2\tau_1^2\tau_2^2\sum_{l=1}^{L^T}\sigma_l^2.
\end{aligned}
\tag{35}
$$

Dividing Eq. (35) both sides by $\frac{\eta\tau_1\tau_2}{2}$ and rearranging terms yields

$$
\begin{aligned}
\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 \leq & \frac{2}{\eta\tau_1\tau_2}\left(\mathbb{E}\left[f(\mathbf{x}^k)\right]-\mathbb{E}\left[f(\mathbf{x}^{k+1})\right]\right)-\sum_{l=1}^{L^T} G_l^2+L^2\eta^2 C\left(\tau_1,\tau_2\right)\sum_{l=1}^{L^T}\left(G_l^2+\sigma_l^2\right)+L\eta\tau_1\tau_2\sum_{l=1}^{L^T} G_l^2 \\
& +L^2\eta^2 D\left(\tau_1,\tau_2\right)\sum_{l=1}^{L_c}\left(G_l^2+\sigma_l^2\right)+L\eta\tau_1\tau_2\sum_{l=1}^{L^T}\sigma_l^2
\end{aligned} \tag{36}
$$

Summing over $k=\{1,2,\ldots,K\}$ and dividing both sides by K yields

$$
\begin{aligned}
& \frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2 \\
\leq & \frac{2}{\eta\tau_1\tau_2 K}\left(f(\mathbf{x}^1)-\mathbb{E}\left[f(\mathbf{x}^{K+1})\right]\right)-\sum_{l=1}^{L^T} G_l^2+L^2\eta^2 C\left(\tau_1,\tau_2\right)\sum_{l=1}^{L^T}\left(G_l^2+\sigma_l^2\right)+L\eta\tau_1\tau_2\sum_{l=1}^{L^T} G_l^2 \\
& +L^2\eta^2 D\left(\tau_1,\tau_2\right)\sum_{l=1}^{L_c}\left(G_l^2+\sigma_l^2\right)+L\eta\tau_1\tau_2\sum_{l=1}^{L^T}\sigma_l^2 \\
& \leq \frac{2}{\eta\tau_1\tau_2 K}\left(f(\mathbf{x}^1)-f^*\right)-\sum_{l=1}^{L^T} G_l^2+L^2\eta^2 C\left(\tau_1,\tau_2\right)\sum_{l=1}^{L^T}\left(G_l^2+\sigma_l^2\right)+L\eta\tau_1\tau_2\sum_{l=1}^{L^T}\left(G_l^2+\sigma_l^2\right) \\
& +L^2\eta^2 D\left(\tau_1,\tau_2\right)\sum_{l=1}^{L_c}\left(G_l^2+\sigma_l^2\right)
\end{aligned} \tag{37}
$$