A PROOF FOR LEMMA 4.2

Proof.

$$
\begin{aligned}
\mathcal{L}_{(t+1)E+0} &= \mathcal{L}_{(t+1)E} + \mathcal{L}_{(t+1)E+0} - \mathcal{L}_{(t+1)E} \\
&\overset{(a)}{=} \mathcal{L}_{(t+1)E} + \mathcal{L}\left(\left(\varphi_k^{t+1}, \theta^{t+1}\right); \boldsymbol{x}, y\right) - \mathcal{L}\left(\left(\varphi_k^{t+1}, \theta_k^{t+1}\right); \boldsymbol{x}, y\right) \\
&\overset{(b)}{\leqslant} \mathcal{L}_{(t+1)E} + \left\langle \nabla\mathcal{L}\left(\left(\varphi_k^{t+1}, \theta_k^{t+1}\right)\right), \left(\left(\varphi_k^{t+1}, \theta^{t+1}\right) - \left(\varphi_k^{t+1}, \theta_k^{t+1}\right)\right)\right\rangle + \frac{L_1}{2}\left\|\left(\varphi_k^{t+1}, \theta^{t+1}\right) - \left(\varphi_k^{t+1}, \theta_k^{t+1}\right)\right\|_2^2 \\
&\overset{(c)}{\leqslant} \mathcal{L}_{(t+1)E} + \frac{L_1}{2}\left\|\left(\varphi_k^{t+1}, \theta^{t+1}\right) - \left(\varphi_k^{t+1}, \theta_k^{t+1}\right)\right\|_2^2 \\
&\overset{(d)}{\leqslant} \mathcal{L}_{(t+1)E} + \frac{L_1}{2}\left\|\theta^{t+1} - \theta_k^{t+1}\right\|_2^2 \\
&\overset{(e)}{=} \mathcal{L}_{(t+1)E} + \frac{L_1}{2}\left\|\theta^{t} - \eta\nabla\mathcal{L}\left(\theta^t\right) - \theta_k^t + \eta\nabla\mathcal{L}\left(\theta_k^t\right)\right\|_2^2 \\
&= \mathcal{L}_{(t+1)E} + \frac{L_1}{2}\left\|\theta^{t} - \theta_k^t + \eta\left(\nabla\mathcal{L}\left(\theta_k^t\right) - \nabla\mathcal{L}\left(\theta^t\right)\right)\right\|_2^2 \\
&\overset{(f)}{\leqslant} \mathcal{L}_{(t+1)E} + \frac{L_1}{2}\left\|\eta\left(\nabla\mathcal{L}\left(\theta_k^t\right) - \nabla\mathcal{L}\left(\theta^t\right)\right)\right\|_2^2 \\
&= \mathcal{L}_{(t+1)E} + \frac{\eta L_1}{2}\left\|\left(\nabla\mathcal{L}\left(\theta_k^t\right) - \nabla\mathcal{L}\left(\theta^t\right)\right)\right\|_2^2.
\end{aligned}
\tag{15}
$$

Take the expectation of $\mathcal{B}$ on both sides of Eq. (15), we have:

$$
\begin{aligned}
\mathbb{E}\left[\mathcal{L}_{(t+1)E+0}\right] &\leqslant \mathbb{E}\left[\mathcal{L}_{(t+1)E}\right] + \frac{\eta L_1}{2}\mathbb{E}\left[\left\|\left(\nabla\mathcal{L}\left(\theta_k^t\right) - \nabla\mathcal{L}\left(\theta^t\right)\right)\right\|_2^2\right] \\
&\overset{(g)}{\leqslant} \mathbb{E}\left[\mathcal{L}_{(t+1)E}\right] + \frac{\eta L_1 \delta^2}{2}.
\end{aligned}
\tag{16}
$$

In Eq. (15), (a): $\mathcal{L}_{(t+1)E+0} = \mathcal{L}\left(\left(\varphi_k^{t+1}, \theta^{t+1}\right); \boldsymbol{x}, y\right)$, i.e., at the start of the $(t+2)$-th round, the k-th client's local model is the combination of the local feature extractor φ_k^{t+1} after local training in the $(t+1)$-th round, and the *global* header θ^{t+1} after training in the $(t+1)$-th round. $\mathcal{L}_{(t+1)E} = \mathcal{L}\left(\left(\varphi_k^{t+1}, \theta_k^{t+1}\right); \boldsymbol{x}, y\right)$, i.e., in the E-th (last) local iteration of the $(t+1)$-th round, the k-th client's local model consists of the feature extractor φ_k^{t+1} and the *local* prediction header θ_k^{t+1}. (b) follows Assumption 4.1. (c): the inequality still holds when the second term is removed from the right-hand side. (d): both $\left(\varphi_k^{t+1}, \theta^{t+1}\right)$ and $\left(\varphi_k^{t+1}, \theta_k^{t+1}\right)$ have the same φ_k^{t+1}, the inequality still holds after it is removed. (e): model training through gradient descent, i.e., $\theta^{t+1} = \theta^t - \eta\nabla\mathcal{L}\left(\theta^t\right), \theta_k^{t+1} = \theta_k^t - \eta\nabla\mathcal{L}\left(\theta_k^t\right)$. Here, we assume that both the learning rate for training local models and the learning rate for training the global prediction header are η. (f): the inequality still holds after removing $\left\|\theta^t - \theta_k^t\right\|_2^2$ from the right hand side. (g) follows Assumption 4.3.

□

B PROOF FOR THEOREM 1

Proof. Substituting Lemma 4.1 into the second term on the right hand side of Lemma 4.2, can have:

$$
\begin{aligned}
\mathbb{E}\left[\mathcal{L}_{(t+1)E+0}\right] &\leqslant \mathcal{L}_{tE+0} - \left(\eta - \frac{L_1\eta^2}{2}\right)\sum_{e=0}^{E}\left\|\mathcal{L}_{tE+e}\right\|_2^2 + \frac{L_1 E\eta^2}{2}\sigma^2 + \frac{\eta L_1\delta^2}{2} \\
&\leqslant \mathcal{L}_{tE+0} - \left(\eta - \frac{L_1\eta^2}{2}\right)\sum_{e=0}^{E}\left\|\mathcal{L}_{tE+e}\right\|_2^2 + \frac{\eta L_1\left(E\eta\sigma^2 + \delta^2\right)}{2}
\end{aligned}
\tag{17}
$$

□

C PROOF FOR THEOREM 2

Proof. Theorem 1 can be re-expressed as:

$$
\sum_{e=0}^{E}\left\|\mathcal{L}_{tE+e}\right\|_2^2 \leqslant \frac{\mathcal{L}_{tE+0} - \mathbb{E}\left[\mathcal{L}_{(t+1)E+0}\right] + \frac{\eta L_1\left(E\eta\sigma^2+\delta^2\right)}{2}}{\eta - \frac{L_1\eta^2}{2}}.
\tag{18}
$$

Take expectations of model ω on both sides of Eq. (18), we have:

$$\sum_{e=0}^{E} \mathbb{E}\left[\|\mathcal{L}_{tE+e}\|_2^2\right] \leqslant \frac{\mathbb{E}\left[\mathcal{L}_{tE+0}\right] - \mathbb{E}\left[\mathcal{L}_{(t+1)E+0}\right] + \frac{\eta L_1(E\eta\sigma^2+\delta^2)}{2}}{\eta - \frac{L_1\eta^2}{2}}. \tag{19}$$

Summing both sides of Eq. (19) over T rounds (i.e., $t \in [0, T-1]$) yields:

$$\frac{1}{T}\sum_{t=0}^{T-1}\sum_{e=0}^{E} \mathbb{E}\left[\|\mathcal{L}_{tE+e}\|_2^2\right] \leqslant \frac{\frac{1}{T}\sum_{t=0}^{T-1}\left(\mathbb{E}\left[\mathcal{L}_{tE+0}\right] - \mathbb{E}\left[\mathcal{L}_{(t+1)E+0}\right]\right) + \frac{\eta L_1(E\eta\sigma^2+\delta^2)}{2}}{\eta - \frac{L_1\eta^2}{2}}. \tag{20}$$

Since $\sum_{t=0}^{T-1}\left(\mathbb{E}\left[\mathcal{L}_{tE+0}\right] - \mathbb{E}\left[\mathcal{L}_{(t+1)E+0}\right]\right) \leqslant \mathcal{L}_{t=0} - \mathcal{L}^*$, we have:

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{e=0}^{E} \mathbb{E}\left[\|\mathcal{L}_{tE+e}\|_2^2\right] &\leqslant \frac{\frac{1}{T}\left(\mathcal{L}_{t=0} - \mathcal{L}^*\right) + \frac{\eta L_1(E\eta\sigma^2+\delta^2)}{2}}{\eta - \frac{L_1\eta^2}{2}} \\ &= \frac{2\left(\mathcal{L}_{t=0} - \mathcal{L}^*\right) + \eta L_1 T\left(E\eta\sigma^2+\delta^2\right)}{T\left(2\eta - L_1\eta^2\right)} \\ &= \frac{2\left(\mathcal{L}_{t=0} - \mathcal{L}^*\right)}{T\eta\left(2 - L_1\eta\right)} + \frac{L_1\left(E\eta\sigma^2+\delta^2\right)}{2 - L_1\eta}.\end{aligned} \tag{21}$$

If the local model can converge, the above equation satisfies

$$\frac{2\left(\mathcal{L}_{t=0} - \mathcal{L}^*\right)}{\mathrm{T}\eta\left(2 - L_1\eta\right)} + \frac{L_1\left(E\eta\sigma^2+\delta^2\right)}{2 - L_1\eta} \leqslant \epsilon. \tag{22}$$

Then, we can obtain:

$$T \geqslant \frac{2\left(\mathcal{L}_{t=0} - \mathcal{L}^*\right)}{\eta\epsilon\left(2 - L_1\eta\right) - \eta L_1\left(E\eta\sigma^2+\delta^2\right)}. \tag{23}$$

Since $T > 0, \mathcal{L}_{t=0} - \mathcal{L}^* > 0$, we can further derive:

$$\eta\epsilon\left(2 - L_1\eta\right) - \eta L_1\left(E\eta\sigma^2+\delta^2\right) > 0, \tag{24}$$

i.e.,

$$\eta < \frac{2\epsilon - L_1\delta^2}{L_1\left(\epsilon + E\sigma^2\right)}. \tag{25}$$

The right-hand side of Eq. (25) are all constants. Thus, the learning rate η is upper bounded. When η satisfies the above condition, the second term of the right-hand side of Eq. (21) is a constant. It can be observed from the first term of Eq. (21) the non-convex convergence rate satisfies $\epsilon \sim \mathcal{O}\left(\frac{1}{T}\right)$. □