

The Convergence Proof Process for FedSOKD-TFA

No Author Given

No Institute Given

Assumption 1. *Lipschitz Smoothness. Gradients of client i 's local complete heterogeneous model w_i are L_1 - Lipschitzsmooth,*

$$\begin{aligned} \|\nabla \mathcal{L}_i^{t_1}(w_i^{t_1}; x, y) - \nabla \mathcal{L}_i^{t_2}(w_i^{t_2}; x, y)\| &\leq L_1 \|w_i^{t_1} - w_i^{t_2}\|, \\ \forall t_1, t_2 > 0, i \in \{0, 1, \dots, N-1\}, (x, y) \in D_i \end{aligned} \quad (1)$$

The above formulation can be further derived as:

$$\mathcal{L}_i^{t_1} - \mathcal{L}_i^{t_2} \leq \langle \nabla \mathcal{L}_i^{t_2}, (w_i^{t_1} - w_i^{t_2}) \rangle + \frac{L_1}{2} \|w_i^{t_1} - w_i^{t_2}\|_2^2 \quad (2)$$

Assumption 2. *Unbiased Gradient and Bounded Variance. Client i 's random gradient $g_t^{w,i} = \nabla \mathcal{L}_i^t(w_i^t, \mathcal{B}_i^t)$, (\mathcal{B} is a batch of local data) is unbiased,*

$$\mathbb{E}_{\mathcal{B}_i^t \subseteq D_i} [g_{w,i}^t] = \nabla \mathcal{L}_i^t(w_i^t) \quad (3)$$

and the variance of random gradient $g_t^{w,i}$ is bounded by:

$$\mathbb{E}_{\mathcal{B}_i^t \subseteq D_i} [\|\nabla \mathcal{L}_i^t(w_i^t; \mathcal{B}_i^t) - \nabla \mathcal{L}_i^t(w_i^t)\|_2^2] \leq \sigma^2 \quad (4)$$

Assumption 3. *Bounded Parameter Variation. The parameter variations of the homogeneous small feature extractor θ_i^t and θ^t before and after aggregation is bounded as*

$$\|\theta^t - \theta_i^t\| \leq \delta^2 \quad (5)$$

Lemma 1. *There is an upper bound on the loss range of any client's local model w in the t local training round.*

$$\mathbb{E}[\mathcal{L}_t^{E+1}] \leq \mathcal{L}_t^{E+0} + \left(\frac{L_1 \eta^2}{2} - \eta\right) \sum_{e=1}^E \|\nabla \mathcal{L}_t^{E+e}\|_2^2 + \frac{L_1 \eta^2 \sigma^2}{2} \quad (6)$$

Lemma 2.

$$\mathcal{L}_{t+1}^{E+0} = \mathcal{L}_{t+1}^E + \mathcal{L}_{t+1}^{E+0} - \mathcal{L}_{t+1}^E \approx \mathcal{L}_{t+1}^E + \eta \|\theta_{t+1}^{E+0} - \theta_{t+1}^E\|_2^2 \leq \mathcal{L}_{t+1}^E + \eta \delta^2 \quad (7)$$

Based on the above assumptions, We can do a further derivation. For convenience, we write an arbitrary client i 's local model as w , and w can be updated by $w_{t+1} = w_t - \eta \eta_{w,t}$, in the $(t+1)$ round, and following Assumption 2, we can obtain

$$\mathcal{L}_t^{E+1} - \mathcal{L}_t^{E+0} \leq \langle \nabla \mathcal{L}_t^{E+0}, (w_t^{E+1} - w_t^{E+0}) \rangle + \frac{L_1}{2} \|w_t^{E+1} - w_t^{E+0}\|_2^2 \quad (8)$$

$$\mathcal{L}_t^{E+1} \leq \mathcal{L}_t^{E+0} - \eta \langle \nabla \mathcal{L}_t^{E+0}, g_w, t_{E+0} \rangle + \frac{L_1 \eta^2}{2} \|g_w, t_{E+0}\|_2^2. \quad (9)$$

Taking the expectation of both sides of the inequality concerning the random variable $\mathcal{E}_{t_{E+0}}$, we obtain

$$\mathbb{E}[\mathcal{L}_t^{E+1}] \leq \mathcal{L}_t^{E+0} - \eta \mathbb{E}[\langle \nabla \mathcal{L}_t^{E+0}, g_w, t_{E+0} \rangle] + \frac{L_1 \eta^2}{2} \mathbb{E}[\|g_w, t_{E+0}\|_2^2] \quad (10)$$

And then, based on the function 3, 4 from Assumption 2 and $Var(x) = \mathbb{E}[x]^2 - (\mathbb{E}[x])^2$, we can derive that

$$\eta \mathbb{E}[\langle \nabla \mathcal{L}_t^{E+0}, g_w, t_{E+0} \rangle] = \eta \|\nabla \mathcal{L}_t^{E+0}\|_2^2 \quad (11)$$

$$\frac{L_1 \eta^2}{2} \mathbb{E}[\|g_w, t_{E+0}\|_2^2] = \frac{L_1 \eta^2}{2} ((\mathbb{E}[\|g_w, t_{E+0}\|_2^2]) + Var(g_w, t_{E+0})) \quad (12)$$

$$\frac{L_1 \eta^2}{2} ((\mathbb{E}[\|g_w, t_{E+0}\|_2^2]) - \|\nabla \mathcal{L}_t^{E+0}\|_2^2) = \frac{L_1 \eta^2}{2} \|\nabla \mathcal{L}_t^{E+0}\|_2^2 \quad (13)$$

Through function 11, 12, 13, we can get that

$$\mathbb{E}[\mathcal{L}_t^{E+1}] \leq \mathcal{L}_t^{E+0} - \eta \|\nabla \mathcal{L}_t^{E+0}\|_2^2 + \frac{L_1 \eta^2}{2} (\|\nabla \mathcal{L}_t^{E+0}\|_2^2 + \sigma^2) \quad (14)$$

Merge items of the same type as above,

$$\mathbb{E}[\mathcal{L}_t^{E+1}] \leq \mathcal{L}_t^{E+0} + (\frac{L_1 \eta^2}{2} - \eta) \|\nabla \mathcal{L}_t^{E+0}\|_2^2 + \frac{L_1 \eta^2 \sigma^2}{2} \quad (15)$$

This concludes the proof of Lemma 1. Next, We're going to prove Lemma 2.

$$\mathcal{L}_{t+1}^{E+0} = \mathcal{L}_{t+1}^E + \mathcal{L}_{t+1}^{E+0} - \mathcal{L}_{t+1}^E \approx \mathcal{L}_{t+1}^E + \eta \|\theta_{t+1}^{E+0} - \theta_{t+1}^E\|_2^2 \leq \mathcal{L}_{t+1}^E + \eta \delta^2 \quad (16)$$

Next, we are going to prove conclusion. Substituting Lemma 1 into the right side of Lemma 2's inequality, we obtain

$$\mathbb{E}[\mathcal{L}_{t+1}^{E+0}] \leq \mathcal{L}_t^{E+0} + (\frac{L_1 \eta^2}{2} - \eta) \sum_{e=1}^E \|\nabla \mathcal{L}_t^{E+e}\|_2^2 + \frac{L_1 \eta^2 \sigma^2}{2} + \eta \delta^2 \quad (17)$$

By transforming, we can obtain

$$\sum_{e=1}^E \|\nabla \mathcal{L}_t^{E+e}\|_2^2 \leq \frac{\mathcal{L}_t^{E+0} - \mathbb{E}[\mathcal{L}_{t+1}^{E+0}] + \frac{L_1 E \eta^2 \sigma^2}{2} + \eta \delta^2}{\eta - \frac{L_1 \eta^2}{2}} \quad (18)$$

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

$$\frac{1}{T} \sum_{t=0}^{T-1} \sum_{e=0}^{E-1} \|\nabla \mathcal{L}_t^{E+e}\|_2^2 \leq \frac{\frac{1}{T} \sum_{t=0}^{T-1} [\mathcal{L}_t^{E+0} - \mathbb{E}[\mathcal{L}_{t+1}^{E+0}]] + \frac{L_1 E \eta^2 \sigma^2}{2} + \eta \delta^2}{\eta - \frac{L_1 \eta^2}{2}} \quad (19)$$

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

$$\frac{1}{T} \sum_{t=0}^{T-1} \sum_{e=0}^{E-1} \|\nabla \mathcal{L}_t^{E+e}\|_2^2 \leq \frac{\frac{\Delta}{T} + \frac{L_1 E \eta^2 \sigma^2}{2} + \eta \delta^2}{\eta - \frac{L_1 \eta^2}{2}} \quad (20)$$

If the above equation converges to a constant ϵ ,

$$\frac{\frac{\Delta}{T} + \frac{L_1 E \eta^2 \sigma^2}{2} + \eta \delta^2}{\eta - \frac{L_1 \eta^2}{2}} \leq \epsilon \quad (21)$$

then

$$T > \frac{\Delta}{\epsilon(\eta - \frac{L_1 \eta^2}{2}) - \frac{L_1 E \eta^2 \sigma^2}{2} - \eta \delta^2} \quad (22)$$

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

$$\epsilon(\eta - \frac{L_1 \eta^2}{2}) - \frac{L_1 E \eta^2 \sigma^2}{2} - \eta \delta^2 > 0 \quad (23)$$

Solving the above inequality yields

$$\eta < \frac{2(\epsilon - \delta^2)}{L_1(\epsilon + E\sigma^2)} \quad (24)$$

Since $\epsilon, L_1, \delta^2, \sigma^2$ are all constants greater than 0, η has solutions. Therefore, when the learning rate η satisfies the above condition, any client's local complete heterogeneous model can converge.