Proof for "Client-Edge-Cloud Hierarchical Federated Learning"

I. System Settings & Algoritms

There are in total L edge servers indexed by ℓ with disjoint client sets $\{C^\ell\}_{\ell=1}^L$, N clients indexed by i with distributed datasets $\{\mathcal{D}_i^\ell\}_{i=1}^N$, $\mathcal{D}^\ell$ denotes the joint dataset under edge ℓ, $|\cdot|$ denotes the size of dataset. τ_1 is the number of local iterations before the clients upload its weights to the server, τ_2 is the number of aggregations performed on the edge before a global aggregation at the remote cloud server. Then with the HierFAVG algorithm in [1], the local model parameters $\boldsymbol{w}_i^\ell(t)$ evolves in the following way:

$$\boldsymbol{w}_i^\ell(t) = \begin{cases} \boldsymbol{w}_i^\ell(t-1) - \eta_t \nabla F_i^\ell(\boldsymbol{w}_i^\ell(t-1)) & t \mid \tau_1 \neq 0 \\ \frac{\sum_{i\in C^\ell}\left[\boldsymbol{w}_i^\ell(t-1)-\eta_t\nabla F_i^\ell(\boldsymbol{w}_i^\ell(t-1))\right]}{|\mathcal{D}^\ell|} & t \mid \tau_1 = 0 \\ & t \mid \tau_1\tau_2 \neq 0 \\ \frac{\sum_{i=1}^N\left[\boldsymbol{w}_i^\ell(t-1)-\eta_t\nabla F_i^\ell(\boldsymbol{w}_i^\ell(t-1))\right]}{|\mathcal{D}|} & t \mid \tau_1\tau_2 = 0 \end{cases} \tag{1}$$

II. Definitions & Assumptions

The total T iterations of update is divided into K cloud intervals with length $\tau_1\tau_2$ and $K\tau_2$ edge intervals with length τ_1. We use $[p]$ to represent the iteration range $[(p-1)\tau_1, p\tau_1]$, $\{q\}$ to represent the iteration range $[(q-1)\tau_1\tau_2, q\tau_1\tau_2]$, so we have $\{q\} = \cup_p[p], \quad p = (q-1)\tau_2 + 1, (q-1)\tau_2 + 2, \ldots, q\tau_2$.

To facilitate our analysis, we introduce several auxiliary sequences and virtual sequences:

- $F^\ell(\boldsymbol{w})$: Edge loss function

$$F^\ell(\boldsymbol{w}) = \frac{\sum_{i\in C^\ell} F_i(\boldsymbol{w})}{|\mathcal{D}^\ell|} \tag{2}$$

This paper presented at the IEEE International Conference on Communications (ICC), 2020 [1]

- $\bar{\boldsymbol{w}}^{\ell}(t)$: weighted average of all the $\boldsymbol{w}_i^{\ell}(t)$ under edge ℓ

$$\bar{\boldsymbol{w}}^{\ell}(t) = \frac{1}{|\mathcal{D}^{\ell}|} \sum_{i \in C^{\ell}} |\mathcal{D}_i^{\ell}| \boldsymbol{w}_i^{\ell}(t) \tag{3}$$

- $\boldsymbol{w}(t)$: weighted average of all the $\boldsymbol{w}_i^{\ell}(t)$

$$\boldsymbol{w}(t) = \frac{1}{|\mathcal{D}|} \sum_{i=1}^{N} |\mathcal{D}_i^{\ell}| \boldsymbol{w}_i^{\ell}(t) \tag{4}$$

- $\boldsymbol{v}_{[p]}^{\ell}(t)$: Virtual edge centralized gradient descent sequence, defined on $[p]$ and evolves following (5), synchronized to $\bar{\boldsymbol{w}}^{\ell}(t)$ at the beginning of $[p]$

$$\begin{aligned} \boldsymbol{v}_{[p]}^{\ell}((p-1)\tau_1) &= \bar{\boldsymbol{w}}^{\ell}((p-1)\tau_1) \\ \boldsymbol{v}_{[p]}^{\ell}(t) &= \boldsymbol{v}_{[p]}^{\ell}(t) - \eta_t \nabla F^{\ell}(\boldsymbol{v}_{[p]}^{\ell}(t-1)) \end{aligned} \tag{5}$$

- $\tilde{\boldsymbol{v}}_{\{q\}}^{\ell}(t)$: Virtual cloud centralized gradient descent sequence, defined on $\{q\}$ and evolves following (6), synchronized to $\boldsymbol{w}(t)$ at the beginning of $\{q\}$

$$\begin{aligned} \tilde{\boldsymbol{v}}_{\{q\}}^{\ell}((q-1)\tau_1\tau_2) &= \boldsymbol{w}((q-1)\tau_1\tau_2) \\ \tilde{\boldsymbol{v}}_{\{q\}}^{\ell}(t) &= \tilde{\boldsymbol{v}}_{\{q\}}^{\ell}(t-1) - \eta_t \nabla F^{\ell}(\tilde{\boldsymbol{v}}_{\{q\}}^{\ell}(t-1)) \end{aligned} \tag{6}$$

- $\boldsymbol{u}_{\{q\}}(t)$: Virtual edge centralized gradient descent sequence, defined on $\{q\}$ and evolves following (7), synchronized to $\boldsymbol{w}(t)$ at the beginning of $\{q\}$

$$\begin{aligned} \boldsymbol{u}_{\{q\}}((q-1)\tau_1\tau_2) &= \boldsymbol{w}((q-1)\tau_1\tau_2) \\ \boldsymbol{u}_{\{q\}}(t) &= \boldsymbol{u}_{\{q\}}(t-1) - \eta_t \nabla F(\boldsymbol{u}_{\{q\}}(t-1)) \end{aligned} \tag{7}$$

We also have following assumptions on the loss functions.

Definition 1 (Gradient Divergence[2]) For any i and $\boldsymbol{w}$, define the gradient divergence between local loss function and edge loss function δ_i^{ℓ} as the upper bound of $\|\nabla F_i^{\ell}(\boldsymbol{w}) - \nabla F^{\ell}(\boldsymbol{w})\|$; the gradient divergence between edge loss function and global loss function Δ^{ℓ} as the upperbound of $\|\nabla F^{\ell}(\boldsymbol{w}) - \nabla F(\boldsymbol{w})\|$, i.e. ,

$$\|\nabla F_i^{\ell}(\boldsymbol{w}) - \nabla F^{\ell}(\boldsymbol{w})\| \le \delta_i^{\ell}$$

$$\|\nabla F^{\ell}(\boldsymbol{w}) - \nabla F(\boldsymbol{w})\| \le \Delta^{\ell}$$

Define $\delta = \frac{\sum_{i=1}^{N} |\mathcal{D}_i^{\ell}| \delta_i^{\ell}}{|\mathcal{D}|}$, $\Delta = \frac{\sum_{\ell=1}^{L} |\mathcal{D}^{\ell}| \Delta^{\ell}}{|\mathcal{D}|} = \frac{\sum_{i=1}^{N} |\mathcal{D}_i^{\ell}| \Delta^{\ell}}{|\mathcal{D}|}$

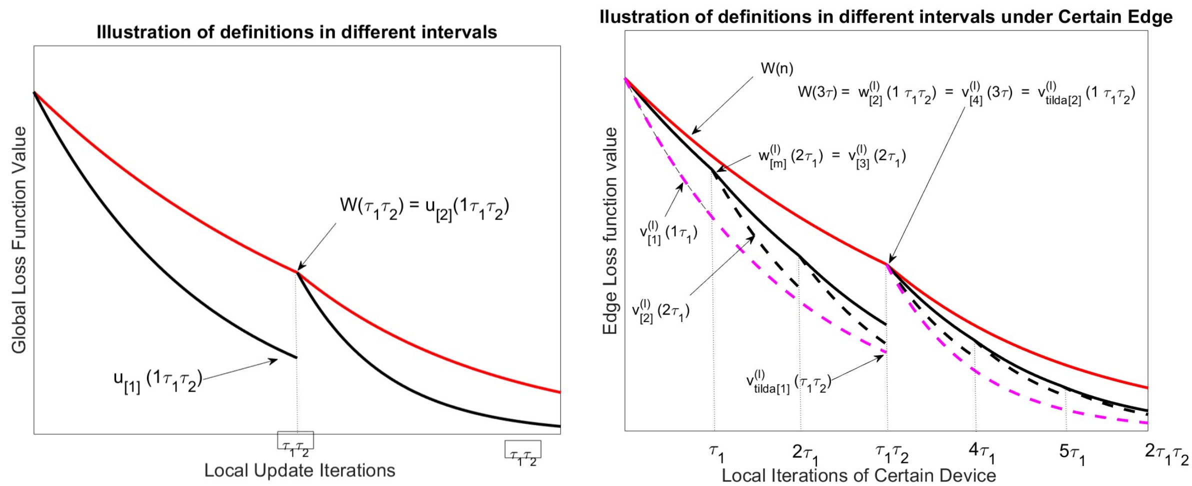

Fig. 1: Illustrations of the virtual sequences and auxiliary variables($\tau_2 = 3$)

Assumption 1 (Lipschitz-continuous) For any i, $F_i^{\ell}(\boldsymbol{w})$ is $\rho-continuous$, i.e., $\|F_i^{\ell}(\boldsymbol{w})-F_i^{\ell}(\boldsymbol{w}')\| \leq \rho\|\boldsymbol{w}-\boldsymbol{w}'\|$

Assumption 2 (Lipschitz-smooth) For any i, $F_i^{\ell}(\boldsymbol{w})$ is $\beta-smooth$, i.e., $\|\nabla F_i^{\ell}(\boldsymbol{w})-\nabla F_i^{\ell}(\boldsymbol{w}')\| \leq \beta\|\boldsymbol{w}-\boldsymbol{w}'\|$

Assumption 3 For any i, $F_i^{\ell}(\boldsymbol{w})$ is convex.

It can be easily inferred that $F^{\ell}(\boldsymbol{w})$ and $F(\boldsymbol{w})$ are all i) $\rho-Lipschitz$ under Assumption 1 ii) $\beta-smooth$ under Assumption 2 iii) convex under Assumption 3

III. Proof of Lemma 2 & Lemma 3 in [1]

Lemma 2 (Convex) With Assumptions 1, 2, and 3, for any cloud aggregation interval $\{q\}$ with a fixed step size η_q, and $t \in \{q\}$, we have

$$\|\boldsymbol{w}(t)-\boldsymbol{u}_{\{q\}}(t)\| \leq G(t,\eta_q) \tag{8}$$

where

$$\begin{aligned} G(t,\eta_q) =& h(t-(q-1)\tau_1\tau_2,\Delta,\eta_q) + h\big(t-((q-1)\tau_2+p(t)-1)\tau_1,\delta,\eta_q\big) \\ &+ \frac{\tau_1}{2}\big(p^2(t)+p(t)-2\big)h(\tau_1,\delta,\eta_q) \end{aligned} \tag{9}$$

$$\begin{aligned} h(x,\delta,\eta) &= \frac{\delta}{\beta}\big((\eta\beta+1)^x-1\big)-\eta\beta x \\ p(x) &= \lceil \frac{x}{\tau_1}-(q-1)\tau_2 \rceil \end{aligned} \tag{10}$$

Proof. From Eqn. (4) and the HierFAVG training algorithm Eqn. (1), we have

$$\boldsymbol{w}(t) = \frac{1}{|\mathcal{D}|}\sum_{i=1}^{N} \boldsymbol{w}_i^{\ell}(t) = \boldsymbol{w}(t-1) - \frac{\eta_q}{|\mathcal{D}|}\sum_{i=1}^{N} \nabla F_i^{\ell}(\boldsymbol{w}_i^{\ell}(t-1)) \tag{11}$$

Thus,

$$\begin{aligned} \|\boldsymbol{w}(t) - \boldsymbol{u}_{\{q\}}(t)\| - \|\boldsymbol{w}(t-1) - \boldsymbol{u}_{\{q\}}(t-1)\| &\leq \frac{\eta_q}{|\mathcal{D}|}\sum_{i=1}^{N} |D_j^{\ell}| \times \|\nabla F_i^{\ell}(w_j^{\ell}(t-1)) - \nabla F_i^{\ell}(u_{\{q\}}t-1))\| \\ &\leq \frac{\eta_q \beta}{|\mathcal{D}|}\sum_{i=1}^{N} |D_j^{\ell}| \times \|w_j^{\ell}(t-1) - u_{\{q\}}(t-1)\|. \end{aligned} \tag{12}$$

Since we have $u_{\{q\}}((q-1)\tau_1\tau_2) = w((q-1)\tau_1\tau_2)$, thus

$$\begin{aligned} \|\boldsymbol{w}(t) - \boldsymbol{u}_{\{q\}}(t)\| = &\sum_{y=(q-1)\tau_1\tau_2+1}^{t} \|\boldsymbol{w}(t) - \boldsymbol{u}_{\{q\}}(t)\| - \|\boldsymbol{w}(t-1) - \boldsymbol{u}_{\{q\}}(t-1)\| \\ \leq &\frac{\eta_q \beta}{|\mathcal{D}|}\sum_{y=(q-1)\tau_1\tau_2+1}^{t}\sum_{i=1}^{N} |D_j^{\ell}| \times \|w_j^{\ell}(y-1) - u_{\{q\}}(y-1)\|. \end{aligned} \tag{13}$$

Next, we will show how to bound $\|w_j^{\ell}(t) - u_{\{q\}}(t)\|$. Using triangle inequality, we have:

$$\|w_j^{\ell}(t) - u_{\{q\}}(t)\| \leq \|w_j^{\ell}(t) - v_{[p]}^{\ell}(t)\| + \|v_{[p]}^{\ell}(t) - \tilde{v}_{\{q\}}^{\ell}(t)\| + \|\tilde{v}_{\{q\}}^{\ell}(t) - u_{\{q\}}(t)\| \tag{14}$$

Using Lemma 3 in [2], we can get:

$$\begin{aligned} &\|w_j^{\ell}(t) - v_{[p]}^{\ell}(t)\| \leq g_i^{\ell}(t-(p-1)\tau_1, \delta_i^{\ell}, \eta_q) \\ &\|\tilde{v}_{\{q\}}^{\ell}(t) - u_{\{q\}}(t)\| \leq g^{\ell}(t-(q-1)\tau_1\tau_2, \Delta^{\ell}, \eta_q) \end{aligned} \tag{15}$$

where function $g(x;\delta)$ is defined as $g_i^{\ell}(x;\delta) = \frac{\delta_i^{\ell}}{\beta}((1+\eta_q\beta)^x - 1)$, function $g^{\ell}(x;\Delta)$ is defined as $g^{\ell}(x;\Delta) = \frac{\Delta^{\ell}}{\beta}((1+\eta_q\beta)^x - 1)$.
By summing it over y and i in Eqn. (13), we have

$$\begin{aligned} \frac{\eta_q \beta}{|\mathcal{D}|}\sum_{y=(q-1)\tau_1\tau_2+1}^{t}\sum_{i=1}^{N} |D_j^{\ell}| \times \|w_j^{\ell}(t) - v_{[p]}^{\ell}(t)\| &\leq \frac{\eta_q \beta}{|\mathcal{D}|}\sum_{y=(q-1)\tau_1\tau_2+1}^{t}\sum_{i=1}^{N} |D_j^{\ell}| \times g_i^{\ell}(t-(p(t)-1)\tau_1, \delta_i^{\ell}, \eta_q) \\ &\leq h(t-((q-1)\tau_2 + p(t) - 2)\tau_1, \delta, \eta_q) \end{aligned} \tag{16}$$

$$\begin{aligned} \frac{\eta_q \beta}{|\mathcal{D}|}\sum_{y=(q-1)\tau_1\tau_2+1}^{t}\sum_{i=1}^{N} |D_j^{\ell}| \times \|\tilde{v}_{\{q\}}^{\ell}(t) - u_{\{q\}}(t)\| &\leq \frac{\eta_q \beta}{|\mathcal{D}|}\sum_{y=(q-1)\tau_1\tau_2+1}^{t}\sum_{i=1}^{N} |D_j^{\ell}| \times g^{\ell}(t-(q-1)\tau_1\tau_2, \Delta^{\ell}, \eta_q) \\ &\leq h(t-(q-1)\tau_1\tau_2, \Delta, \eta_q) \end{aligned} \tag{17}$$

where $p(t) = \lceil \frac{t}{\tau_1} - (q-1)\tau_2 \rceil$, $p(t)$ is defined the index of edge aggregation in the cloud interval $\{q\}$.

For $\|v_{[p]}^{\ell}(t) - \tilde{v}_{\{q\}}^{\ell}(t)\|$, we will show that it is bounded by a step function and increases at $t = (q-1)\tau_1\tau_2 + p\tau_1, p = 1, \ldots, \tau_2 - 1$, which means that for $t \in [p]$,

$$\|v_{[p]}^{\ell}(t) - \tilde{v}_{\{q\}}^{\ell}(t)\| \leq \|v_{[p]}^{\ell}((p-1)\tau_1) - \tilde{v}_{\{q\}}^{\ell}((p-1)\tau_1)\| \tag{18}$$

From Eqn. (5) and (6), we can see that when $t \in [p]$ $v_{[p]}^{\ell}(t)$ and $\tilde{v}_{\{q\}}^{\ell}(t)$, the evolution of these two parameters are equal to performing gradient descent on the same loss function from different initialization point. A convex and smooth function $f(x)$ satisfies the following:

$$\begin{aligned} \langle \nabla f(x) - \nabla f(y), x - y \rangle &\geq \mu \|x - y\|_2^2, \\ \langle \nabla f(x) - \nabla f(y), x - y \rangle &\geq \frac{1}{\beta} \|\nabla f(x)) - \nabla f(y)\|_2^2 \end{aligned} \tag{19}$$

where $\mu \geq 0$. Thus, it satisfies the regularity condition, i.e.,

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{1}{2}\mu \|x - y\|_2^2 + \frac{1}{2\beta} \|\nabla f(x)) - \nabla f(y)\|_2^2 \tag{20}$$

Thus for $v_{[p]}^{\ell}(t+1)$ and $\tilde{v}_{\{q\}}^{\ell}(t+1)$,

$$\begin{aligned} &\left\| v_{[p]}^{\ell}(t+1) - \tilde{v}_{\{q\}}^{\ell}(t+1) \right\|_2^2 - \left\| v_{[p]}^{\ell}(t) - \tilde{v}_{\{q\}}^{\ell}(t) \right\|_2^2 \\ =& \left\| \nabla F^{\ell}(v_{[p]}^{\ell}(t)) - \nabla F^{\ell} \tilde{v}_{\{q\}}^{\ell}(t)) \right\|_2^2 - 2\eta \langle v_{[p]}^{\ell}(t) - \tilde{v}_{\{q\}}^{\ell}(t), \nabla F^{\ell}(v_{[p]}^{\ell}(t)) - \nabla F^{\ell}(\tilde{v}_{\{q\}}^{\ell}(t)) \rangle \\ \leq& 0 \end{aligned} \tag{21}$$

when $\eta \leq \frac{1}{\beta}$. Thus, Eqn. (18) is proved. The incremental value of the step function is $h(\tau_1, \delta, \eta_q)$. Thus,

$$\|v_{[p]}^{\ell}(t) - \tilde{v}_{\{q\}}^{\ell}(t)\| \leq p(t) * h(\tau_1, \delta, \eta_q) \tag{22}$$

summing Eqn. 22, we have

$$\frac{\eta_q \beta}{|\mathcal{D}|} \sum_{y=(q-1)\tau_1\tau_2+1}^{t} \sum_{i=1}^{N} |D_j^{\ell}| \times \|v_{[p]}^{\ell}(t) - \tilde{v}_{\{q\}}^{\ell}(t)\| \leq \frac{\tau_1}{2} \big(p^2(t) + p(t) - 2\big) h(\tau_1, \delta, \eta_q) \tag{23}$$

By adding the bounded value of the three terms in Eqn. (14), the bound in Lemma 2 is derived.

□

Lemma 3 (Non-convex) For any i, assuming $f_i(w)$ is β-smooth, for any cloud interval $\{q\}$ with step size η_q, we have

$$\|\boldsymbol{w}(t) - \boldsymbol{u}_{\{q\}}(t)\| \leq G_{nc}(\kappa_1\kappa_2, \eta_q)$$

where

$$G_{nc}(\kappa_1\kappa_2, \eta_q) = h(\kappa_1\kappa_2, \Delta, \eta_q) + \kappa_1\kappa_2\frac{(1+\eta_q\beta)^{\kappa_1\kappa_2} - 1}{(1+\eta_q\beta)^{\kappa_1} - 1}h(\kappa_1, \delta, \eta_q) + h(\kappa_1, \delta, \eta_q),$$

$$h(x, \delta, \eta) = \frac{\delta}{\beta}\big((\eta\beta + 1)^x - 1\big) - \eta\beta x.$$

Proof. The proof process is the same as the convex case except for bounding the $\|v^{\ell}_{[p]}(t) - \tilde{v}^{\ell}_{\{q\}}(t)\|$, since the regularity condition is nor satisfied for non-convex functions. For non-convex loss functions, we have

$$\begin{aligned}
\|v^{\ell}_{[p]}(t+1) - \tilde{v}^{\ell}_{\{q\}}(t+1)\| &= \|v^{\ell}_{[p]}(t) - \tilde{v}^{\ell}_{\{q\}}(t) - (\eta_q \nabla F^{\ell}(v^{\ell}_{[p]}(t+1)) - \eta_q \nabla F^{\ell}\tilde{v}^{\ell}_{\{q\}}(t+1))\| \\
&\leq \|v^{\ell}_{[p]}(t) - \tilde{v}^{\ell}_{\{q\}}(t)\| + \eta_q\beta\|v^{\ell}_{[p]}(t) - \tilde{v}^{\ell}_{\{q\}}(t)\| \\
&\leq (1+\eta_q\beta)\|v^{\ell}_{[p]}(t) - \tilde{v}^{\ell}_{\{q\}}(t)\| \\
&\leq (1+\eta_q\beta)^{t-(p-1)\tau_1}\|v^{\ell}_{[p]}((p-1)\tau_1) - \tilde{v}^{\ell}_{\{q\}}((p-1)\tau_1)\|
\end{aligned} \tag{24}$$

$\square$

REFERENCES

[1] L. Liu, J. Zhang, S. H. Song, and K. B. Letaief, “Client-Edge-Cloud Hierarchical Federated Learning,” in *IEEE International Conference on Communications (ICC)*, 2020, pp. 1–6.

[2] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, “When edge meets learning: Adaptive control for resource-constrained distributed machine learning,” in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 2018, pp. 63–71.