Supplemental Materials to Subspace-divided Federated Representation Learning

ANONYMOUS AUTHORS

ACM Reference Format:

Anonymous Authors. 2022. Supplemental Materials to Subspace-divided Federated Representation Learning. In . ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 PROOF OF THEOREM 1

1.1 Mathematical Notations

Let $\mathcal{G}_t$ be the local objective function of client t, a.k.a, the model in the client. Then FDRL aims to

$$arg\min_{\mathbf{W}_t} \mathcal{G}_t = arg\min_{\mathbf{W}_t} \sum_{i=1}^{n_t} \mathcal{L}_t(\mathcal{F}_t(\mathcal{M}_{\widetilde{\mathbf{W}}^t}(\mathbf{x}_i), \mathcal{M}_{\widehat{\mathbf{W}}^t}(\mathbf{x}_i))) + \alpha \mathcal{S}_m(\mathcal{M}_{\widetilde{\mathbf{W}}^t}, \mathcal{M}_{\widehat{\mathbf{W}}^t}) + \beta \mathcal{S}_r(\mathcal{M}_{\widetilde{\mathbf{W}}^t}, \mathcal{M}_{\widehat{\mathbf{W}}^t}) \tag{1}$$

where $\mathcal{M}_{\widetilde{\mathbf{W}}^t}$ and $\mathcal{M}_{\widehat{\mathbf{W}}^t}$ are the shared model and the not-shared model in client t, $\mathcal{F}_t(\cdot)$ is the integration function, $\mathcal{S}_m(\cdot)$ is the model distance function and $\mathcal{S}_r(\cdot)$ is the representation distance function. Both $\mathcal{S}_m(\cdot)$ and $\mathcal{S}_r(\cdot)$ are convex functions.

TO BE CONVENIENT in proofs, we consider FDRL sharing the entire model but just returning the same parameters expect for the shared sub-model. Let $\mathcal{G}$ be the global objective function, that is

$$arg\min_{\mathbf{W}^*} \mathcal{G} = arg\min_{\mathbf{W}^*} \sum_{t=1}^{m} \mathcal{G}_{\mathbf{W}^*}(\mathcal{G}_t) \tag{2}$$

where $\mathbf{W}^*$ is the parameter of the global model.

For simplicity, we let $\mathbf{W}_k^t$, $\mathbf{Q}_k^t$ and $\mathbf{H}_k^t$ be the parameters of the shared sub-model, the not-shared sub-model and the local head in client t at e k-th step, where $\mathbf{Q}_k^t$ and $\mathbf{H}_k^t$ are abused by

$$\begin{aligned} \mathbf{Q}_k^t &= [\mathbf{0}, \cdots, \frac{1}{p_t}\mathbf{Q}_k^t, \cdots, \mathbf{0}]^T \in \mathcal{R}^m \\ \mathbf{H}_k^t &= [\mathbf{0}, \cdots, \frac{1}{p_t}\mathbf{H}_k^t, \cdots, \mathbf{0}]^T \in \mathcal{R}^m. \end{aligned} \tag{3}$$

Besides, we denote by $\mathbf{Q}_k = [\mathbf{Q}_k^1, \mathbf{Q}_k^2, \cdots, \mathbf{Q}_k^m]^T$ and $\mathbf{H}_k = [\mathbf{H}_k^1, \mathbf{H}_k^2, \cdots, \mathbf{H}_k^m]^T$ the parameters of all sub-models and heads in m clients, respectively.

Let I_τ be the set of update steps of the FDRL model, i.e., $I_\tau = \{n\tau \mid n = 1, 2, \cdots\}$. If $k+1 \notin I_\tau$, each local client updates parameters by stochastic gradient descent (*sgd*). If $k+1 \in I_\tau$, all clients communicate with the server. The parameters in the client are updated by

$$(\mathbf{W'}_{k+1}^t, \mathbf{Q'}_{k+1}^t, \mathbf{H'}_{k+1}^t) = (\mathbf{W}_k^t, \mathbf{Q}_k^t, \mathbf{H}_k^t) - \eta_k \nabla \mathcal{G}_t(\mathbf{W}_k^t, \mathbf{Q}_k^t, \mathbf{H}_k^t; \xi_k^t) \tag{4}$$

$$(\mathbf{W}_{k+1}^t, \mathbf{Q}_{k+1}^t, \mathbf{H}_{k+1}^t) = \begin{cases} (\mathbf{W'}_{k+1}^t, \mathbf{Q'}_{k+1}^t, \mathbf{H'}_{k+1}^t) & \text{if } k+1 \notin I_\tau \\ (\sum_{t=1}^m p_t \mathbf{W'}_{k+1}^t, \mathbf{Q'}_{k+1}^t, \mathbf{H'}_{k+1}^t) & \text{if } k+1 \in I_\tau \end{cases} \tag{5}$$

where m is the number of clients; η_k is the learning rate on k-th step; $p_t = 1/m$ is the weight for the t-th client.

In Eq. (4), $(\mathbf{W'}_{k+1}^t, \mathbf{Q'}_{k+1}^t, \mathbf{H'}_{k+1}^t)$ represents the updated result of the *sgd* step from $(\mathbf{W}_k^t, \mathbf{Q}_k^t, \mathbf{H}_k^t)$. And in Eq. (5), when $k+1 \in I_\tau$, one aggregation step is performed to update $\mathbf{W}_{k+1}^t$, and when $k+1 \notin I_\tau$, $(\mathbf{W}_{k+1}^t, \mathbf{Q}_{k+1}^t, \mathbf{H}_{k+1}^t)$ is still equal to $(\mathbf{W'}_{k+1}^t, \mathbf{Q'}_{k+1}^t, \mathbf{H'}_{k+1}^t)$. To have an intuition on this setting, we can write the update link into

$$(\mathbf{W}_0^t, \mathbf{Q}_0^t, \mathbf{H}_0^t) \xrightarrow{sgd} (\mathbf{W'}_1^t, \mathbf{Q'}_1^t, \mathbf{H'}_1^t) \xrightarrow{:=} (\mathbf{W}_1^t, \mathbf{Q}_1^t, \mathbf{H}_1^t) \xrightarrow{sgd} (\mathbf{W'}_2^t, \mathbf{Q'}_2^t, \mathbf{H'}_2^t) \xrightarrow{:=} \cdots \xrightarrow{sgd} (\mathbf{W'}_\tau^t, \mathbf{Q'}_\tau^t, \mathbf{H'}_\tau^t) \xrightarrow{comm}$$
$$(\mathbf{W}_\tau^t, \mathbf{Q}_\tau^t, \mathbf{H}_\tau^t) \xrightarrow{sgd} (\mathbf{W'}_{\tau+1}^t, \mathbf{Q'}_{\tau+1}^t, \mathbf{H'}_{\tau+1}^t) \xrightarrow{:=} (\mathbf{W}_{\tau+1}^t, \mathbf{Q}_{\tau+1}^t, \mathbf{H}_{\tau+1}^t) \xrightarrow{sgd} \cdots \xrightarrow{sgd} (\mathbf{W'}_{2\tau}^t, \mathbf{Q'}_{2\tau}^t, \mathbf{H'}_{2\tau}^t) \xrightarrow{comm} \cdots$$

From this update link, we define two virtual sequences as follows:

$$\begin{aligned} (\overline{\mathbf{W}}'_k, \mathbf{Q}'_k, \mathbf{H}'_k) &= \sum_{t=1}^m p_t (\mathbf{W'}_k^t, \mathbf{Q'}_k^t, \mathbf{H'}_k^t) = (\sum_{t=1}^m p_t \mathbf{W'}^t, \mathbf{Q}'_k, \mathbf{H}'_k) \\ (\overline{\mathbf{W}}_k, \mathbf{Q}_k, \mathbf{H}_k) &= \sum_{t=1}^m p_t (\mathbf{W}_k^t, \mathbf{Q}_k^t, \mathbf{H}_k^t) = (\sum_{t=1}^m p_t \mathbf{W}_k^t, \mathbf{Q}_k, \mathbf{H}_k) \end{aligned} \tag{6}$$

where $(\overline{\mathbf{W}}'_{k+1}, \mathbf{Q}'_{k+1}, \mathbf{H}'_{k+1})$ is the actual result of one SGD step from $(\overline{\mathbf{W}}_k, \mathbf{Q}_k, \mathbf{H}_k)$.

In addition, we let $\overline{\mathbf{g}}_k = \sum_{t=1}^m p_t \nabla \mathcal{G}_t(\mathbf{W}_k^t, \mathbf{Q}_k^t, \mathbf{H}_k^t)$ and $\mathbf{g}_k = \sum_{t=1}^m p_t \nabla \mathcal{G}_t(\mathbf{W}_k^t, \mathbf{Q}_k^t, \mathbf{H}_k^t; \xi_k^t)$, leading to $\mathbb{E}\mathbf{g}_k = \overline{\mathbf{g}}_k$.

With these notations, we perform the average operator on two sides of Eq. (4) and then arrive at

$$(\overline{\mathbf{W}}'_{k+1}, \mathbf{Q}'_{k+1}, \mathbf{H}'_{k+1}) = (\overline{\mathbf{W}}_k, \mathbf{Q}_k, \mathbf{H}_k) - \eta_k \mathbf{g}_k.$$

1.2 Lemmas

To analysis the convergence of FDRL, we use the proof method proposed in the work of Li et al. [1] in our study. Frist of all, we set up the following three lemmas that can be proven easily by using the prior works. Note that we here rewrite them for self-contained explanations. Please see the literature [1] for more details.

Lemma 1.1. *Let Assumptions 1 and 2 hold. If $\eta_k \le \frac{1}{4L}$, we have*

$$\begin{aligned} \mathbb{E}\|(\overline{\mathbf{W}}'_{k+1}, \mathbf{Q}'_{k+1}, \mathbf{H}'_{k+1}) - (\mathbf{W}^+, \mathbf{Q}^+, \mathbf{H}^+)\|^2 &\le (1-\eta_k\mu)\mathbb{E}\|(\overline{\mathbf{W}}_k, \mathbf{Q}_k, \mathbf{H}_k) - (\mathbf{W}^+, \mathbf{Q}^+, \mathbf{H}^+)\|^2 + \eta_k^2 \mathbb{E}\|\mathbf{g}_k - \overline{\mathbf{g}}_k\|^2 \\ &\quad + 6L\eta_k^2\Gamma + 2\mathbb{E}\sum_{t=1}^m p_t \|(\overline{\mathbf{W}}_k, \mathbf{Q}_k, \mathbf{H}_k) - (\mathbf{W}_k^t, \mathbf{Q}_k^t, \mathbf{H}_k^t)\|^2. \end{aligned}$$

where $(\mathbf{W}^+, \mathbf{Q}^+, \mathbf{H}^+)$ is the optimal parameter, $\Gamma = \mathcal{G}^+ - \sum_{t=1}^m p_t \mathcal{G}_t^+ \ge 0$, $\mathcal{G}^+$ is the minimum value of global update and $\mathcal{G}_t^+$ is the minimum value of $\mathcal{G}_t$.

Lemma 1.2. *Let Assumption 3 hold.*

$$\mathbb{E}\|\mathbf{g}_k - \overline{\mathbf{g}}_k\|^2 \le \sum_{t=1}^m p_t^2 \sigma_t^2.$$

Lemma 1.3. *Let Assumption 4 hold. η_k is non-increasing and $\eta_k \le 2\eta_{k+\tau}$ for all $t \ge 0$.*

$$\mathbb{E}\left[\sum_{t=1}^m p_t \|(\overline{\mathbf{W}}_k, \mathbf{Q}_k, \mathbf{H}_k) - (\mathbf{W}_k^t, \mathbf{Q}_k^t, \mathbf{H}_k^t)\|^2\right] \le 4\eta_k^2(\tau-1)^2G^2.$$

1.3 Proof of Theorem 1

Proof. From the previous definitions in Eq. (4), we have the following two analyses.

1) If $k+1 \notin I_\tau$,

$$(\overline{\mathbf{W}}_{k+1}, \mathbf{Q}_{k+1}, \mathbf{H}_{k+1}) = \sum_{t=1}^{m} p_t (\mathbf{W}_{k+1}^t, \mathbf{Q}_{k+1}^t, \mathbf{H}_{k+1}^t) = \sum_{t=1}^{m} p_t (\mathbf{W'}_{k+1}^{t}, \mathbf{Q'}_{k+1}^{t}, \mathbf{H'}_{k+1}^{t}) = (\overline{\mathbf{W}}'_{k+1}, \mathbf{Q}'_{k+1}, \mathbf{H}'_{k+1})$$

2) If $k+1 \in I_\tau$,

$$\begin{aligned}(\overline{\mathbf{W}}_{k+1}, \mathbf{Q}_{k+1}, \mathbf{H}_{k+1}) &= \sum_{t=1}^{m} p_t (\mathbf{W}_{k+1}^t, \mathbf{Q}_{k+1}^t, \mathbf{H}_{k+1}^t) = \sum_{t=1}^{m} p_t (\sum_{t=1}^{m} p_t \mathbf{W'}_{k+1}^{t}, \mathbf{Q'}_{k+1}^{t}, \mathbf{H'}_{k+1}^{t}) \\ &= (\sum_{t=1}^{m} p_t \sum_{t=1}^{m} p_t \mathbf{W'}_{k+1}^{t}, \sum_{t=1}^{m} p_t \mathbf{Q'}_{k+1}^{t}, \sum_{t=1}^{m} p_t \mathbf{H'}_{k+1}^{t}) \\ &= (\sum_{t=1}^{m} p_t \overline{\mathbf{W}}'_{k+1}, \mathbf{Q}'_{k+1}, \mathbf{H}'_{k+1}) \\ &= (\overline{\mathbf{W}}'_{k+1}, \mathbf{Q}'_{k+1}, \mathbf{H}'_{k+1})\end{aligned}$$

Hence, $(\overline{\mathbf{W}}_{k+1}, \mathbf{Q}_{k+1}, \mathbf{H}_{k+1}) = (\overline{\mathbf{W}}'_{k+1}, \mathbf{Q}'_{k+1}, \mathbf{H}'_{k+1})$. Denote by $\Delta_k = \mathbb{E}\|(\overline{\mathbf{W}}_k, \mathbf{Q}_k, \mathbf{H}_k) - (\mathbf{W}^+, \mathbf{Q}^+, \mathbf{H}^+)\|^2$. From Lemma1, Lemma 2, Lemma 3, like that used in [1], we also arrive at

$$\Delta_{k+1} \le (1 - \eta_k \mu)\Delta_k + \eta_k^2 B \tag{7}$$

where

$$B = \sum_{t=1}^{m} p_t^2 \sigma_t^2 + 6L\Gamma + 8(\tau - 1)^2 G^2.$$

Then, referring to the prior work [1], we let $\kappa = L/\mu$, $\gamma = max\{8\kappa, \tau\}$ and $\eta_k = \frac{2}{\mu(\gamma+k)}$, and finally have

$$\mathbb{E}[\mathcal{G}(\overline{\mathbf{W}}_k, \mathbf{Q}_k, \mathbf{H}_k)] - \mathcal{G}^+ \le \frac{\kappa}{\gamma + K - 1}(\frac{2B}{\mu} + \frac{\mu\gamma}{2}\mathbb{E}\|(\overline{\mathbf{W}}_1, \mathbf{Q}_1, \mathbf{H}_1) - (\mathbf{W}^+, \mathbf{Q}^+, \mathbf{H}^+)\|^2) \tag{8}$$

where

$$B = \sum_{t=1}^{m} p_t^2 \sigma_t^2 + 6L\Gamma + 8(\tau - 1)^2 G^2.$$

□

2 PROOF OF THEOREM 2

2.1 Mathematical Notations

With respect to the case of partial client participation in FDRL, a random multiset S_k will be chosen in the k-th communication round. In our setting, S_k contains a subset of T indices uniformly sampled from all clients without replacement. That is, each client in S_k only occurs once in a communication round. We also consider our problem as a problem of sharing all parameters but keeping the local non-shared sub-model and the local head, and then rewrite the update link as

$$(\mathbf{W'}_{k+1}^{t}, \mathbf{Q'}_{k+1}^{t}, \mathbf{H'}_{k+1}^{t}) = (\mathbf{W}_k^t, \mathbf{Q}_k^t, \mathbf{H}_k^t) - \eta_k \nabla \mathcal{G}_t(\mathbf{W}_k^t, \mathbf{Q}_k^t, \mathbf{H}_k^t; \xi_k^t) \tag{9}$$

$$(\mathbf{W}_{k+1}^{t},\mathbf{Q}_{k+1}^{t},\mathbf{H}_{k+1}^{t})=\begin{cases}(\mathbf{W}_{k+1}^{\prime\,t},\mathbf{Q}_{k+1}^{\prime\,t},\mathbf{H}_{k+1}^{\prime\,t}) & \text{if } k+1\notin I_\tau\\ (\sum\limits_{t\in S_{k+1}} p_t\frac{m}{T}\mathbf{W}_{k+1}^{\prime\,t},\mathbf{Q}_{k+1}^{\prime\,t},\mathbf{H}_{k+1}^{\prime\,t}) & \text{if } k+1\in I_\tau\end{cases} \tag{10}$$

where m is the number of clients; η_k is the learning rate on k-th step; $p_t = 1/m$ in our study is the weight for the t-th client.

2.2 Lemmas

Lemma 2.1. *If $k+1 \in I_\tau$, there has*

$$\mathbb{E}_{S_k}(\overline{\mathbf{W}}_{k+1},\mathbf{Q}_{k+1},\mathbf{H}_{k+1})=(\overline{\mathbf{W}}_{k+1}^{\prime},\mathbf{Q}_{k+1}^{\prime},\mathbf{H}_{k+1}^{\prime}). \tag{11}$$

Proof. In [1], there has an important observation. That can be expressed as follows. Letting $\{x_i\}_{i=1}^{m}$ denote a fixed sequence, a multiset S_k with size T is sampled in the k-th round, where x_t is sampled with probability q_t. Let $S_k = \{i_1, .., i_T\} \subset [m]$. The observation [1] now is

$$\mathbb{E}_{S_k}\sum_{t\in S_k}x_t=\mathbb{E}_{S_k}\sum_{t=1}^{T}x_{i_t}=T\mathbb{E}_{S_k}x_{i_1}=T\sum_{t=1}^{m}q_t x_t$$

where $q_t = \frac{1}{m}$ in our study, then

$$\begin{aligned}\mathbb{E}_{S_k}(\overline{\mathbf{W}}_{k+1},\mathbf{Q}_{k+1},\mathbf{H}_{k+1}) &= \mathbb{E}_{S_k}\sum_{t=1}^{m}p_t(\mathbf{W}_{k+1}^{t},\mathbf{Q}_{k+1}^{t},\mathbf{H}_{k+1}^{t})\\ &=\sum_{t=1}^{m}p_t\mathbb{E}_{S_k}(\sum_{t\in S_{k+1}}p_t\frac{m}{T}\mathbf{W}_{k+1}^{\prime\,t},\mathbf{Q}_{k+1}^{\prime\,t},\mathbf{H}_{k+1}^{\prime\,t})\\ &=\sum_{t=1}^{m}p_t(\mathbb{E}_{S_k}(\sum_{t\in S_{k+1}}p_t\frac{m}{T}\mathbf{W}_{k+1}^{\prime\,t}),\mathbb{E}_{S_k}(\mathbf{Q}_{k+1}^{\prime\,t}),\mathbb{E}_{S_k}(\mathbf{H}_{k+1}^{\prime\,t}))\\ &=\sum_{t=1}^{m}p_t(p_t\frac{m}{T}\mathbb{E}_{S_k}(\sum_{t\in S_{k+1}}\mathbf{W}_{k+1}^{\prime\,t}),\mathbb{E}_{S_k}(\mathbf{Q}_{k+1}^{\prime\,t}),\mathbb{E}_{S_k}(\mathbf{H}_{k+1}^{\prime\,t}))\\ &=\sum_{t=1}^{m}p_t(p_t\frac{m}{T}T\sum_{t=1}^{m}q_t\mathbf{W}_{k+1}^{\prime\,t},\sum_{t=1}^{m}q_t\mathbf{Q}_{k+1}^{\prime\,t},\sum_{t=1}^{m}q_t\mathbf{H}_{k+1}^{\prime\,t})\\ &=(\sum_{t=1}^{m}p_tp_t\frac{m}{T}T\sum_{t=1}^{m}q_t\mathbf{W}_{k+1}^{\prime\,t},\sum_{t=1}^{m}p_t\sum_{t=1}^{m}q_t\mathbf{Q}_{k+1}^{\prime\,t},\sum_{t=1}^{m}p_t\sum_{t=1}^{m}q_t\mathbf{H}_{k+1}^{\prime\,t})\\ &=(\sum_{t=1}^{m}q_t\overline{\mathbf{W}}_{k+1}^{\prime},\sum_{t=1}^{m}q_t\mathbf{Q}_{k+1}^{\prime},\sum_{t=1}^{m}q_t\mathbf{H}_{k+1}^{\prime})\\ &=(\overline{\mathbf{W}}_{k+1}^{\prime},\mathbf{Q}_{k+1}^{\prime},\mathbf{H}_{k+1}^{\prime})\end{aligned}$$

Thus, the conclusion in Eq. (11) is obtained. □

Lemma 2.2. *If $k+1 \in I_\tau$, assume that η_k is non-increasing and $\eta_k \le 2\eta_{k+\tau}$ for all $k \ge 0$. With $p_t = 1/m$, the expected difference between $(\overline{\mathbf{W}}_{k+1}^{\prime},\mathbf{Q}_{k+1}^{\prime},\mathbf{H}_{k+1}^{\prime})$ and $(\overline{\mathbf{W}}_{k+1},\mathbf{Q}_{k+1},\mathbf{H}_{k+1})$ is bounded by*

$$\mathbb{E}_{S_k}\|(\overline{\mathbf{W}}_{k+1}^{\prime},\mathbf{Q}_{k+1}^{\prime},\mathbf{H}_{k+1}^{\prime})-(\overline{\mathbf{W}}_{k+1},\mathbf{Q}_{k+1},\mathbf{H}_{k+1})\|^2\le\frac{m-T}{m-1}\frac{4}{T}\eta_k^2\tau^2G^2$$

Proof. It is easy to reach by using the prior works in [1]. Please refer for more details. □

2.3 Proof of Theorem 2

Proof. Towards the convergence analysis, we have

$$\begin{aligned} &\|(\overline{\mathbf{W}}_{k+1}, \mathbf{Q}_{k+1}, \mathbf{H}_{k+1}) - (\mathbf{W}^+, \mathbf{Q}^+, \mathbf{H}^+)\|^2 \\ &= \|(\overline{\mathbf{W}}_{k+1}, \mathbf{Q}_{k+1}, \mathbf{H}_{k+1}) - (\overline{\mathbf{W}}'_{k+1}, \mathbf{Q}'_{k+1}, \mathbf{H}'_{k+1}) + (\overline{\mathbf{W}}'_{k+1}, \mathbf{Q}'_{k+1}, \mathbf{H}'_{k+1}) - (\mathbf{W}^+, \mathbf{Q}^+, \mathbf{H}^+)\|^2 \\ &= \|(\overline{\mathbf{W}}_{k+1}, \mathbf{Q}_{k+1}, \mathbf{H}_{k+1}) - (\overline{\mathbf{W}}'_{k+1}, \mathbf{Q}'_{k+1}, \mathbf{H}'_{k+1})\|^2 + \|(\overline{\mathbf{W}}'_{k+1}, \mathbf{Q}'_{k+1}, \mathbf{H}'_{k+1}) - (\mathbf{W}^+, \mathbf{Q}^+, \mathbf{H}^+)\|^2 \\ &+ 2\langle(\overline{\mathbf{W}}_{k+1}, \mathbf{Q}_{k+1}, \mathbf{H}_{k+1}) - (\overline{\mathbf{W}}'_{k+1}, \mathbf{Q}'_{k+1}, \mathbf{H}'_{k+1}), (\overline{\mathbf{W}}'_{k+1}, \mathbf{Q}'_{k+1}, \mathbf{H}'_{k+1}) - (\mathbf{W}^+, \mathbf{Q}^+, \mathbf{H}^+)\rangle. \end{aligned} \tag{12}$$

where the third term will vanishes resulted from Lemma 4 if taking expectation over S_{k+1}.
1) If $k+1 \notin I_\tau$, it is

$$(\overline{\mathbf{W}}_{k+1}, \mathbf{Q}_{k+1}, \mathbf{H}_{k+1}) = \sum_{t=1}^{m} p_t(\mathbf{W}^t_{k+1}, \mathbf{Q}^t_{k+1}, \mathbf{H}^t_{k+1}) = \sum_{t=1}^{m} p_t(\mathbf{W}'^{\,t}_{k+1}, \mathbf{Q}'^{\,t}_{k+1}, \mathbf{H}'^{\,t}_{k+1}) = (\overline{\mathbf{W}}'_{k+1}, \mathbf{Q}'_{k+1}, \mathbf{H}'_{k+1}). \tag{13}$$

Thus, the first term in Eq. (12), i.e., $\|(\overline{\mathbf{W}}_{k+1}, \mathbf{Q}_{k+1}, \mathbf{H}_{k+1}) - (\overline{\mathbf{W}}'_{k+1}, \mathbf{Q}'_{k+1}, \mathbf{H}'_{k+1})\|^2 = 0$. And then, following Lemma 1, Lemma 2 and Lemma 3, we have

$$\mathbb{E}\|(\overline{\mathbf{W}}_{k+1}, \mathbf{Q}_{k+1}, \mathbf{H}_{k+1}) - (\mathbf{W}^+, \mathbf{Q}^+, \mathbf{H}^+)\|^2 \le (1-\eta_k\mu)\mathbb{E}\|(\overline{\mathbf{W}}_k, \mathbf{Q}_k, \mathbf{H}_k) - (\mathbf{W}^+, \mathbf{Q}^+, \mathbf{H}^+)\|^2 + \eta_k^2 B$$

2) If $k+1 \in I_\tau$, we use Lemma 5,

$$\mathbb{E}\|(\overline{\mathbf{W}}_{k+1}, \mathbf{Q}_{k+1}, \mathbf{H}_{k+1}) - (\mathbf{W}^+, \mathbf{Q}^+, \mathbf{H}^+)\|^2 \le (1-\eta_k\mu)\mathbb{E}\|(\overline{\mathbf{W}}_k, \mathbf{Q}_k, \mathbf{H}_k) - (\mathbf{W}^+, \mathbf{Q}^+, \mathbf{H}^+)\|^2 + \eta_k^2 (B+C)$$

Finally, referring to the prior works [1], we let κ, γ, η_k and B be defined in Theorem 1 and achieve

$$\mathbb{E}[\mathcal{G}(\overline{\mathbf{W}}_k, \mathbf{Q}_k, \mathbf{H}_k)] - \mathcal{G}^+ \le \frac{\kappa}{\gamma+K-1}\left(\frac{2(B+C)}{\mu} + \frac{\mu\gamma}{2}\mathbb{E}\|(\overline{\mathbf{W}}_1, \mathbf{Q}_1, \mathbf{H}_1) - (\mathbf{W}^+, \mathbf{Q}^+, \mathbf{H}^+)\|^2\right) \tag{14}$$

where $C = \frac{4(m-T)}{T(m-1)}\tau^2 G^2$. □

REFERENCES

[1] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. 2019. On the Convergence of FedAvg on Non-IID Data. In *International Conference on Learning Representations*.