Supplementary Material

Yining Wang, Wanli Ni, Wenqiang Yi, *Member, IEEE,*
Xiaodong Xu, *Senior Member, IEEE,* Ping Zhang, *Fellow, IEEE,*
and Arumugam Nallanathan, *Fellow, IEEE*

I. Preliminaries

Preliminary 1. The loss function for each client k is defined as

$$\mathcal{L}_k = \frac{1}{D_k} \sum_{c=1}^{C} \sum_{i \in \mathcal{D}_{k,c}} \mathcal{L}_T(r_{k,i}, y_{k,i}) + \lambda \| \hat{\boldsymbol{f}}_{k,i} - \overline{\boldsymbol{F}}^c \|_2^2, \tag{1}$$

where D_k denotes the number of data samples from client k, c denotes the semantic concept, $\mathcal{D}_{k,c}$ denotes the set of data samples belonging to semantic c from client k. $\mathcal{L}_T(r_{k,i}, y_{k,i})$ denotes the task loss, where $r_{k,i}$ is the predicted logits of sample i and $y_{k,i}$ is the label. $\hat{\boldsymbol{f}}_{k,i}$ is the noised semantic feature of sample i, and $\overline{\boldsymbol{F}}^c$ denotes the global semantic centroid of concept c. λ is the regularization coefficient.

Preliminary 2. Each personalized semantic encoder is trained on client k and is parameterized by $\boldsymbol{\theta}_k$, each parallel semantic decoder is trained on the BS server and is parameterized by $\boldsymbol{\phi}_k$. Therefore, the entire parameter set of client k is denoted as $\boldsymbol{w}_k = \{\boldsymbol{\theta}_k, \boldsymbol{\phi}_k\}$.

Yining Wang is with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, 100876, China (e-mail: joanna_wyn@bupt.edu.cn).

Wanli Ni is with Department of Electronic Engineering, Tsinghua University, Beijing, 100084, China (e-mail: niwanli@tsinghua.edu.cn).

Wenqiang Yi is with the School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, U.K. (e-mail: wy23627@essex.ac.uk).

Xiaodong Xu and Ping Zhang are with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China, and also with Peng Cheng Laboratory, Shenzhen 518066, China (e-mail: xuxiaodong@bupt.edu.cn; pzhang@bupt.edu.cn).

Arumugam Nallanathan is with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, U.K. (e-mail: a.nallanathan@qmul.ac.uk).

Preliminary 3. The global semantic centroids are generated as

$$\overline{\boldsymbol{F}} = \sigma(\boldsymbol{\varphi}), \tag{2}$$

where $\overline{\boldsymbol{F}} = \{\overline{\boldsymbol{F}}^c\}_{c=1}^C$ and $\sigma(\cdot)$ represents the SCG network.

II. Proof of Theorem 1

To make a proof of Theorem 1, we first derive the following lemmas:

Lemma 1: Let Assumptions 1 and 2 hold. From the beginning of communication round $t+1$ to the last local update step, the loss function of an arbitrary client can be bounded as:

$$\mathbb{E}[\mathcal{L}_{(t+1)E}] \leq \mathcal{L}_{tE+1/2} - (\eta - \frac{L_1\eta^2}{2}) \sum_{e=1/2}^{E-1} \|\nabla\mathcal{L}_{tE+e}\|_2^2 + \frac{L_1 E\eta^2}{2}\rho^2. \tag{3}$$

Proof: Since this lemma is valid for an arbitrary client, the client notation k is omitted. Let $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \boldsymbol{g}_t$, then

$$\begin{aligned} \mathcal{L}_{tE+1} &\overset{(a)}{\leq} \mathcal{L}_{tE+1/2} + \langle \nabla\mathcal{L}_{tE+1/2}, (\boldsymbol{w}_{tE+1} - \boldsymbol{w}_{tE+1/2})\rangle + \frac{L_1}{2}\|\boldsymbol{w}_{tE+1} - \boldsymbol{w}_{tE+1/2}\|_2^2 \\ &= \mathcal{L}_{tE+1/2} - \eta\langle \nabla\mathcal{L}_{tE+1/2}, \boldsymbol{g}_{tE+1/2}\rangle + \frac{L_1}{2}\|\eta \boldsymbol{g}_{tE+1/2}\|_2^2, \end{aligned} \tag{4}$$

where (a) follows from the quadratic L_1-Lipschitz smooth bound in Assumption 1. Taking expectation of both sides of the above equation on the random variable $\xi_{tE+1/2}$, we have

$$\begin{aligned} \mathbb{E}[\mathcal{L}_{tE+1}] &\leq \mathcal{L}_{tE+1/2} - \eta\mathbb{E}[\langle \nabla\mathcal{L}_{tE+1/2}, \boldsymbol{g}_{tE+1/2}\rangle] + \frac{L_1\eta^2}{2}\mathbb{E}[\|\boldsymbol{g}_{tE+1/2}\|_2^2] \\ &\overset{(b)}{=} \mathcal{L}_{tE+1/2} - \eta\|\nabla\mathcal{L}_{tE+1/2}\|_2^2 + \frac{L_1\eta^2}{2}\mathbb{E}[\|\boldsymbol{g}_{k,tE+1/2}\|_2^2] \\ &\overset{(c)}{\leq} \mathcal{L}_{tE+1/2} - \eta\|\nabla\mathcal{L}_{tE+1/2}\|_2^2 + \frac{L_1\eta^2}{2}(\|\nabla\mathcal{L}_{tE+1/2}\|_2^2 + Var(\boldsymbol{g}_{k,tE+1/2})) \\ &= \mathcal{L}_{tE+1/2} - (\eta - \frac{L_1\eta^2}{2})\|\nabla\mathcal{L}_{tE+1/2}\|_2^2 + \frac{L_1\eta^2}{2}Var(\boldsymbol{g}_{k,tE+1/2}) \\ &\overset{(d)}{\leq} \mathcal{L}_{tE+1/2} - (\eta - \frac{L_1\eta^2}{2})\|\nabla\mathcal{L}_{tE+1/2}\|_2^2 + \frac{L_1\eta^2}{2}\rho^2, \end{aligned} \tag{5}$$

where (b) follows from Assumption 2, (c) follows from $Var(x) = \mathbb{E}[x^2] - (\mathbb{E}[x])^2$, (d) follows from Assumption 2. Take expectation of $\boldsymbol{w}$ on both sides and telescope E steps, we have,

$$\mathbb{E}[\mathcal{L}_{(t+1)E}] \leq \mathcal{L}_{tE+1/2} - (\eta - \frac{L_1\eta^2}{2}) \sum_{e=1/2}^{E-1} \|\nabla\mathcal{L}_{tE+e}\|_2^2 + \frac{L_1 E\eta^2}{2}\rho^2. \tag{6}$$

■

Lemma 2: Let Assumptions 3 and 4 hold. After the global semantic centroid generating at the BS server, the loss function of an arbitrary client can be bounded as:

$$\mathbb{E}[\mathcal{L}_{(t+1)E+1/2}] \leq \mathcal{L}_{(t+1)E} + \lambda L_2 \eta E' G. \tag{7}$$

Proof:

$$\begin{aligned}
\mathcal{L}_{(t+1)E+1/2} &= \mathcal{L}_{(t+1)E} + \mathcal{L}_{(t+1)E+1/2} - \mathcal{L}_{(t+1)E} \\
&\overset{(a)}{=} \mathcal{L}_{(t+1)E} + \lambda \|\alpha(\boldsymbol{\theta}_{(t+1)E}) - \overline{\boldsymbol{F}}_{t+2}\|_2 - \lambda \|\alpha(\boldsymbol{\theta}_{(t+1)E}) - \overline{\boldsymbol{F}}_{t+1}\|_2 \\
&\overset{(b)}{\leq} \mathcal{L}_{(t+1)E} + \lambda \|\overline{\boldsymbol{F}}_{t+2} - \overline{\boldsymbol{F}}_{t+1}\|_2 \\
&\overset{(c)}{=} \mathcal{L}_{(t+1)E} + \lambda \|\sigma(\boldsymbol{\varphi}_{(t+2)E'}) - \sigma(\boldsymbol{\varphi}_{(t+1)E'})\|_2 \\
&\overset{(d)}{\leq} \mathcal{L}_{(t+1)E} + \lambda L_2 \|\boldsymbol{\varphi}_{(t+2)E'} - \boldsymbol{\varphi}_{(t+1)E'}\|_2 \\
&= \mathcal{L}_{(t+1)E} + \lambda L_2 \eta \|\sum_{e=0}^{E'-1} \boldsymbol{g}'_{(t+1)E'+e}\|_2 \\
&\overset{(e)}{\leq} \mathcal{L}_{(t+1)E} + \lambda L_2 \eta \sum_{e=0}^{E'-1} \|\boldsymbol{g}'_{(t+1)E'+e}\|_2.
\end{aligned} \tag{8}$$

Take expectations of random variable ξ on both sides, then

$$\begin{aligned}
\mathbb{E}[\mathcal{L}_{(t+1)E+1/2}] &\leq \mathcal{L}_{(t+1)E} + \lambda L_2 \eta \sum_{e=0}^{E'-1} \mathbb{E}[\|\boldsymbol{g}'_{(t+1)E'+e}\|_2] \\
&\overset{(f)}{\leq} \mathcal{L}_{(t+1)E} + \lambda L_2 \eta E' G,
\end{aligned} \tag{9}$$

where (a) follows from the definition of local loss function in (1), (b) follows from $\|a-b\|_2 - \|a-c\|_2 \leq \|b-c\|_2$, (c) follows from the definition of global semantic centroids in (2), (d) follows from L_2-Lipschitz continuity in Assumption 3, (e) follows from $\|\sum a_i\|_2 \leq \sum \|a_i\|_2$, and (f) follows from Assumption 4. ■

Taking expectation of $\boldsymbol{w}$ on both sides of Lemma 1 and 2, then sum them together, we have

$$\mathbb{E}[\mathcal{L}_{(t+1)E+1/2}] \leq \mathcal{L}_{tE+1/2} - (\eta - \frac{L_1 \eta^2}{2}) \sum_{e=1/2}^{E-1} \|\nabla \mathcal{L}_{tE+e}\|_2^2 + \frac{L_1 E \eta^2}{2} \rho^2 + \lambda L_2 \eta E' G. \tag{10}$$

Thus, Theorem 1 is proven.

III. Proof of Corollary 1

Corollary 1: (Non-convex FedCL convergence). The loss function $\mathcal{L}$ of an arbitrary client monotonously decreases in every communication round when

$$\eta_{e'} < \frac{2\left(\sum_{e=1/2}^{e'} \|\nabla \mathcal{L}_{tE+e}\|_2^2 - \lambda L_2 E' G\right)}{L_1 \left(\sum_{e=1/2}^{e'} \|\nabla \mathcal{L}_{tE+e}\|_2^2 + E\rho^2\right)}, \quad e' = \frac{1}{2}, 1, ..., E-1, \tag{11}$$

and

$$\lambda_t < \frac{\|\nabla \mathcal{L}_{tE+e}\|_2^2}{L_2 E' G}. \tag{12}$$

Thus, the loss function converges.

Proof: As observed in Theorem 1, to guarantee certain one-round decrease, it satisfies $-(\eta - \frac{L_1\eta^2}{2}) \sum_{e=1/2}^{E-1} \|\nabla \mathcal{L}_{tE+e}\|_2^2 + \frac{L_1 E \eta^2}{2}\rho^2 + \lambda L_2 \eta E' G \leq 0$, therefore we have

$$\eta < \frac{2\left(\sum_{e=1/2}^{E-1} \|\nabla \mathcal{L}_{tE+e}\|_2^2 - \lambda L_2 E' G\right)}{L_1 \left(\sum_{e=1/2}^{E-1} \|\nabla \mathcal{L}_{tE+e}\|_2^2 + E\rho^2\right)}, \tag{13}$$

and

$$\lambda < \frac{\sum_{e=1/2}^{E-1} \|\nabla \mathcal{L}_{tE+e}\|_2^2}{L_2 E' G}. \tag{14}$$

Following the practical set up in FedProto, we use

$$\eta_{e'} < \frac{2\left(\sum_{e=1/2}^{e'} \|\nabla \mathcal{L}_{tE+e}\|_2^2 - \lambda L_2 E' G\right)}{L_1 \left(\sum_{e=1/2}^{e'} \|\nabla \mathcal{L}_{tE+e}\|_2^2 + E\rho^2\right)}, \quad e' = \frac{1}{2}, 1, ..., E-1, \tag{15}$$

and

$$\lambda_t < \frac{\|\nabla \mathcal{L}_{tE+e}\|_2^2}{L_2 E' G}. \tag{16}$$

Thus, the convergence of $\mathcal{L}$ holds, which proves Corollary 1. ∎

IV. Proof of Theorem 2

Theorem 2: (Non-convex convergence rate of FedCL). Let Assumptions 1 to 4 hold and $\Delta = \mathcal{L}_0 - \mathcal{L}^*$, for an arbitrary client, given any $\epsilon > 0$, after

$$T = \frac{2\Delta}{E\epsilon(2\eta - L_1\eta^2) - E\eta^2\rho^2 L_1 - 2\lambda\eta L_2 E' G} \tag{17}$$

communication rounds of the FedCL framework, we have

$$\frac{1}{TE} \sum_{t=0}^{T-1} \sum_{e=1/2}^{E-1} \mathbb{E}[\|\nabla \mathcal{L}_{tE+e}\|_2^2] < \epsilon, \tag{18}$$

if

$$\eta < \frac{2(E\epsilon - \lambda L_2 E' G)}{L_1 E(\epsilon + \rho^2)}, \tag{19}$$

and

$$\lambda < \frac{E\epsilon}{L_2 E' G}. \tag{20}$$

Proof: Take expectation of $\boldsymbol{w}$ on both sides of Theorem 1, and then telescope from round $t = 0$ to $t = T - 1$ with step from $e = \frac{1}{2}$ to $e = E$ in each round, we have

$$\begin{aligned} &\frac{1}{TE}\sum_{t=0}^{T-1}\sum_{e=1/2}^{E-1}\mathbb{E}[\|\nabla\mathcal{L}_{tE+e}\|_2^2] \\ \leq &\frac{\frac{1}{TE}\sum_{t=0}^{T-1}\left(\mathcal{L}_{tE+1/2} - \mathbb{E}[\mathcal{L}_{(t+1)E+1/2}]\right) + \frac{L_1\eta^2}{2}\rho^2 + \frac{\lambda L_2 \eta E' G}{E}}{\eta - \frac{L_1\eta^2}{2}}. \end{aligned} \tag{21}$$

Given any $\epsilon > 0$, let the right term $< \epsilon$, we have

$$\frac{\frac{2}{TE}\sum_{t=0}^{T-1}\left(\mathcal{L}_{tE+1/2} - \mathbb{E}[\mathcal{L}_{(t+1)E+1/2}]\right) + L_1\eta^2\rho^2 + \frac{2\lambda L_2\eta E' G}{E}}{2\eta - L_1\eta^2} < \epsilon. \tag{22}$$

Let $\Delta = \mathcal{L}_0 - \mathcal{L}^*$, since $\sum_{t=0}^{T-1}\left(\mathcal{L}_{tE+1/2} - \mathbb{E}[\mathcal{L}_{(t+1)E+1/2}]\right) \leq \Delta$, the above formulation (22) is valid when

$$\frac{\frac{2\Delta}{TE} + L_1\eta^2\rho^2 + \frac{2\lambda L_2\eta E' G}{E}}{2\eta - L_1\eta^2} < \epsilon, \tag{23}$$

that is,

$$T > \frac{2\Delta}{E\epsilon(2\eta - L_1\eta^2) - E\eta^2\rho^2 L_1 - 2\lambda\eta L_2 E' G}. \tag{24}$$

Therefore, it is proven that

$$\frac{1}{TE}\sum_{t=0}^{T-1}\sum_{e=1/2}^{E-1}\mathbb{E}[\|\nabla\mathcal{L}_{tE+e}\|_2^2] < \epsilon, \tag{25}$$

when

$$\eta < \frac{2(E\epsilon - \lambda L_2 E' G)}{L_1 E(\epsilon + \rho^2)}, \tag{26}$$

and

$$\lambda < \frac{E\epsilon}{L_2 E' G}. \tag{27}$$

■