Supplementary Material for the Paper: "Retransmission-Based Semi-Federated Learning"

Jingheng Zheng, Hui Tian, Wanli Ni, Gaofeng Nie, Wenchao Jiang, and Tony Q. S. Quek, *Fellow, IEEE*

In the document, we present the derivation of Theorem 1 in detail.

Appendix C
Proof of Theorem 1

Based on (9), it is derived that

$$\begin{aligned}\mathbf{w}_{t+1}-\mathbf{w}_t =&\hat{\rho}_t(\Delta\hat{\mathbf{w}}_t-\Delta\hat{\mathbf{w}}_t')+\hat{\rho}_t\Delta\hat{\mathbf{w}}_t'+\tilde{\rho}_t\Delta\tilde{\mathbf{w}}_t\\ =&\hat{\rho}_t(\Delta\hat{\mathbf{w}}_t-\Delta\hat{\mathbf{w}}_t')\\ &-\eta\hat{\rho}_t\left(\sum\nolimits_{k=1}^{K}q_{t,k}\sum\nolimits_{i=1}^{I_t}\nabla\hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,i-1})\right)\\ &-\eta\tilde{\rho}_t\left(\sum\nolimits_{i=1}^{I_t}\nabla\tilde{F}_t(\tilde{\mathbf{w}}_{t,i-1})\right). \qquad \text{(A.1)}\end{aligned}$$

By plugging $\mathbf{w}=\mathbf{w}_{t+1}$ and $\mathbf{w}'=\mathbf{w}_t$ as well as (A.1) into (27), we have

$$\begin{aligned}&F(\mathbf{w}_{t+1})-F(\mathbf{w}_t)\\ \leq&\hat{\rho}_t\nabla F(\mathbf{w}_t)^{\mathrm{T}}(\Delta\hat{\mathbf{w}}_t-\Delta\hat{\mathbf{w}}_t')\\ &-\eta\hat{\rho}_t\sum\nolimits_{k=1}^{K}q_{t,k}\sum\nolimits_{i=1}^{I_t}\nabla F(\mathbf{w})^{\mathrm{T}}\nabla\hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,i-1})\\ &-\eta\tilde{\rho}_t\sum\nolimits_{i=1}^{I_t}\nabla F(\mathbf{w}_t)^{\mathrm{T}}\nabla\tilde{F}_t(\tilde{\mathbf{w}}_{t,i-1})\\ &+\frac{L}{2}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|^2,\\ \overset{(a)}{=}&\hat{\rho}_t\nabla F(\mathbf{w}_t)^{\mathrm{T}}(\Delta\hat{\mathbf{w}}_t-\Delta\hat{\mathbf{w}}_t')-\frac{\eta I_t}{2}\|\nabla F(\mathbf{w}_t)\|^2\\ &-\frac{\eta\hat{\rho}_t}{2}\sum\nolimits_{k=1}^{K}q_{t,k}\sum\nolimits_{i=1}^{I_t}\|\nabla\hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,i-1})\|^2\\ &-\frac{\eta\tilde{\rho}_t}{2}\sum\nolimits_{i=1}^{I_t}\|\nabla\tilde{F}_t(\tilde{\mathbf{w}}_{t,i-1})\|^2\\ &+\frac{\eta\hat{\rho}_t}{2}\sum\nolimits_{k=1}^{K}q_{t,k}\sum\nolimits_{i=1}^{I_t}\|\nabla F(\mathbf{w}_t)-\nabla\hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,i-1})\|^2\\ &+\frac{\eta\tilde{\rho}_t}{2}\sum\nolimits_{i=1}^{I_t}\|\nabla F(\mathbf{w}_t)-\nabla\tilde{F}_t(\tilde{\mathbf{w}}_{t,i-1})\|^2\\ &+\frac{L}{2}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|^2\\ \overset{(b)}{\leq}&\hat{\rho}_t\nabla F(\mathbf{w}_t)^{\mathrm{T}}(\Delta\hat{\mathbf{w}}_t-\Delta\hat{\mathbf{w}}_t')-\frac{\eta I_t}{2}\|\nabla F(\mathbf{w}_t)\|^2\\ &-\frac{\eta\hat{\rho}_t}{2}\sum\nolimits_{k=1}^{K}q_{t,k}\sum\nolimits_{i=1}^{I_t}\|\nabla\hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,i-1})\|^2\\ &-\frac{\eta\tilde{\rho}_t}{2}\sum\nolimits_{i=1}^{I_t}\|\nabla\tilde{F}_t(\tilde{\mathbf{w}}_{t,i-1})\|^2\\ &+\frac{3\eta\hat{\rho}_t}{2}\sum\nolimits_{k=1}^{K}q_{t,k}\sum\nolimits_{i=1}^{I_t}\Big[\|\nabla F(\mathbf{w}_t)-\nabla\hat{F}_{t,k}(\mathbf{w}_t)\|^2\\ &+\|\nabla\hat{F}_{t,k}(\mathbf{w}_t)\|^2+\|\nabla\hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,i-1})\|^2\Big]\\ &+\frac{3\eta\tilde{\rho}_t}{2}\sum\nolimits_{i=1}^{I_t}\Big[\|\nabla F(\mathbf{w}_t)-\nabla\tilde{F}_t(\mathbf{w}_t)\|^2\\ &+\|\nabla\tilde{F}_t(\mathbf{w}_t)\|^2+\|\nabla\tilde{F}_t(\tilde{\mathbf{w}}_{t,i-1})\|^2\Big]\\ &+\frac{L}{2}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|^2\\ \overset{(c)}{\leq}&\hat{\rho}_t\nabla F(\mathbf{w}_t)^{\mathrm{T}}(\Delta\hat{\mathbf{w}}_t-\Delta\hat{\mathbf{w}}_t')-\frac{\eta I_t}{2}\|\nabla F(\mathbf{w}_t)\|^2\\ &-\frac{\eta\hat{\rho}_t}{2}\sum\nolimits_{k=1}^{K}q_{t,k}\sum\nolimits_{i=1}^{I_t}\|\nabla\hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,i-1})\|^2\\ &-\frac{\eta\tilde{\rho}_t}{2}\sum\nolimits_{i=1}^{I_t}\|\nabla\tilde{F}_t(\tilde{\mathbf{w}}_{t,i-1})\|^2\\ &+\frac{3\eta I_t}{2}(\delta^2+2G^2)+\frac{L}{2}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|^2, \qquad \text{(A.2)}\end{aligned}$$

where (a) is because $\mathbf{x}^{\mathrm{T}}\mathbf{y}=\frac{1}{2}(\|\mathbf{x}\|^2+\|\mathbf{y}\|^2-\|\mathbf{x}-\mathbf{y}\|^2), \forall\mathbf{x},\mathbf{y}$, (b) comes from applying the Cauchy-Schwarz inequality, and (c) is due to Assumptions 3 and 4.

We now bound $\|\mathbf{w}_{t+1}-\mathbf{w}_t\|^2$ as follows:

$$\begin{aligned}&\|\mathbf{w}_{t+1}-\mathbf{w}_t\|^2\\ \leq&2\hat{\rho}_t^2\|\Delta\hat{\mathbf{w}}_t-\Delta\hat{\mathbf{w}}_t'\|^2\\ &+2\eta^2\Bigg\|\sum\nolimits_{k=1}^{K}\sum\nolimits_{i=1}^{I_t}\hat{\rho}_t q_{t,k}\nabla\hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,i-1})\\ &+\sum\nolimits_{i=1}^{I_t}\tilde{\rho}_t\nabla\tilde{F}_t(\tilde{\mathbf{w}}_{t,i-1})\Bigg\|^2\\ \overset{(d)}{\leq}&2\hat{\rho}_t^2\|\Delta\hat{\mathbf{w}}_t-\Delta\hat{\mathbf{w}}_t'\|^2+2\eta^2\left(\sum\nolimits_{k=1}^{K}\sum\nolimits_{i=1}^{I_t}\hat{\rho}_t q_{t,k}+\sum\nolimits_{i=1}^{I_t}\tilde{\rho}_t\right)\\ &\times\Bigg(\sum\nolimits_{k=1}^{K}\sum\nolimits_{i=1}^{I_t}\hat{\rho}_t q_{t,k}\|\nabla\hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,i-1})\|^2\\ &+\sum\nolimits_{i=1}^{I_t}\tilde{\rho}_t\|\nabla\tilde{F}_t(\tilde{\mathbf{w}}_{t,i-1})\|^2\Bigg)\\ \overset{(e)}{\leq}&2\hat{\rho}_t\|\Delta\hat{\mathbf{w}}_t-\Delta\hat{\mathbf{w}}_t'\|^2+2\eta^2 I_t\tilde{\rho}_t\sum\nolimits_{i=1}^{I_t}\|\nabla\tilde{F}_t(\tilde{\mathbf{w}}_{t,i-1})\|^2\\ &+2\eta^2 I_t\hat{\rho}_t\sum\nolimits_{k=1}^{K}q_{t,k}\sum\nolimits_{i=1}^{I_t}\|\nabla\hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,i-1})\|^2, \qquad \text{(A.3)}\end{aligned}$$

where (d) is because of the Cauchy-Schwarz inequality and

Jingheng Zheng, Hui Tian and Gaofeng Nie are with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: zhengjh@bupt.edu.cn; tianhui@bupt.edu.cn; niegaofeng@bupt.edu.cn).

Wanli Ni is with the department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: charleswall@bupt.edu.cn).

Wenchao Jiang and Tony Q. S. Quek are with the Information System Technology and Design Pillar, Singapore University of Technology and Design, Singapore 487372. Tony Q. S. Quek is also with the Department of Electronic Engineering, Kyung Hee University, Yongin 17104, South Korea (e-mail: wenchao_jiang@sutd.edu.sg; tonyquek@sutd.edu.sg).

(e) is because $\hat{\rho}_t^2 \leq \hat{\rho}_t$ when $\hat{\rho}_t \in [0,1]$. Plugging (A.3) and $\eta = 1/L$ into (A.2), we have

$$
\begin{aligned}
& F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t) \\
\leq & \hat{\rho}_t \nabla F(\mathbf{w}_t)^{\mathrm{T}} (\Delta\hat{\mathbf{w}}_t - \Delta\hat{\mathbf{w}}_t') - \frac{I_t}{2L}\|\nabla F(\mathbf{w}_t)\|^2 \\
& + \hat{\rho}_t \left(\frac{2I_t-1}{2L}\right) \sum\nolimits_{k=1}^{K} q_{t,k} \sum\nolimits_{i=1}^{I_t} \|\nabla \hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,i-1})\|^2 \\
& + \tilde{\rho}_t \left(\frac{2I_t-1}{2L}\right) \sum\nolimits_{i=1}^{I_t} \|\nabla \tilde{F}_t(\tilde{\mathbf{w}}_{t,i-1})\|^2 \\
& + \frac{3I_t}{2L}(\delta^2 + 2G^2) + L\hat{\rho}_t \|\Delta\hat{\mathbf{w}}_t - \Delta\hat{\mathbf{w}}_t'\|^2 \\
\overset{(f)}{\leq} & \frac{1}{2L}\left[\hat{\rho}_t \left(\frac{L}{\mu}\gamma^{\mathrm{A}} - I_t\right) - \tilde{\rho}_t I_t\right] \|\nabla F(\mathbf{w}_t)\|^2 \\
& + \hat{\rho}_t L \left(\frac{1}{2\gamma^{\mathrm{A}}} + 1\right) \|\Delta\hat{\mathbf{w}}_t - \Delta\hat{\mathbf{w}}_t'\|^2 + \frac{3I_t}{2L}(\delta^2 + 2G^2) \\
& + \hat{\rho}_t \left(\frac{2I_t-1}{2L}\right) \sum\nolimits_{k=1}^{K} q_{t,k} \sum\nolimits_{i=1}^{I_t} \|\nabla \hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,i-1})\|^2 \\
& + \tilde{\rho}_t \left(\frac{2I_t-1}{2L}\right) \sum\nolimits_{i=1}^{I_t} \|\nabla \tilde{F}_t(\tilde{\mathbf{w}}_{t,i-1})\|^2 \\
\overset{(g)}{\leq} & \frac{1}{2L}\left[\hat{\rho}_t \left(\frac{L}{\mu}\gamma^{\mathrm{A}} - I_t\right) - \tilde{\rho}_t I_t\right] \|\nabla F(\mathbf{w}_t)\|^2 + \frac{3\delta^2 + 5G^2}{2L} I_t \\
& + \hat{\rho}_t L \left(\frac{1}{2\gamma^{\mathrm{A}}} + 1\right) \|\Delta\hat{\mathbf{w}}_t - \Delta\hat{\mathbf{w}}_t'\|^2 + \frac{G^2}{L} {I_t}^2,
\end{aligned} \tag{A.4}
$$

where (f) is because $\hat{\rho}_t \nabla F(\mathbf{w})^{\mathrm{T}} (\Delta\hat{\mathbf{w}}_t - \Delta\hat{\mathbf{w}}_t') \leq \hat{\rho}_t[\gamma^{\mathrm{A}}\|\nabla F(\mathbf{w}_t)\|^2/(2L) + L\|\Delta\hat{\mathbf{w}}_t - \Delta\hat{\mathbf{w}}_t'\|^2/(2\gamma^{\mathrm{A}})]$ and $L/\mu \geq 1$, and (g) comes from applying Assumption 4.

Suppose the normalization and de-normalization procedures for AirComp-based aggregation of local model updates which are proposed in our previous work [1] are employed in this paper. Then, one can verify that

$$
\begin{aligned}
\mathbb{E}[\|\Delta\hat{\mathbf{w}}_t - \Delta\hat{\mathbf{w}}_t'\|^2] \leq & \sum\nolimits_{i=1}^{Q^{\mathrm{M}}} \mathbb{E}[\bar{\sigma}_t^2 \mathrm{MSE}_{t,i}^{\mathrm{A}}] \\
\leq & \sum\nolimits_{i=1}^{Q^{\mathrm{M}}} \mathbb{E}[\bar{\sigma}_t^2]\gamma^{\mathrm{A}} = \mathbb{E}[\bar{\sigma}_t^2] Q^{\mathrm{M}} \gamma^{\mathrm{A}},
\end{aligned} \tag{A.5}
$$

where $\bar{\sigma}_t^2$ denotes the global variance of $\Delta\hat{\mathbf{w}}_{t,k} = [\Delta\hat{w}_{t,k,1}, \ldots, \Delta\hat{w}_{t,k,i}, \ldots, \Delta\hat{w}_{t,k,Q^{\mathrm{M}}}]^{\mathrm{T}}, \forall k \in \mathcal{K}$, defined by

$$
\begin{aligned}
\bar{\sigma}_t^2 = & \frac{1}{K} \sum\nolimits_{k=1}^{K} \left(\frac{1}{Q^{\mathrm{M}}} \sum\nolimits_{q=1}^{Q^{\mathrm{M}}} \Delta\hat{w}_{t,k,q}^2\right) \\
& - \left(\frac{1}{K} \sum\nolimits_{k=1}^{K} \frac{1}{Q^{\mathrm{M}}} \sum\nolimits_{q=1}^{Q^{\mathrm{M}}} \Delta\hat{w}_{t,k,q}\right)^2.
\end{aligned} \tag{A.6}
$$

Based on Assumption 4 and (3) as well as $\eta = 1/L$, we bound $\mathbb{E}[\bar{\sigma}_t^2]$ as follows:

$$
\begin{aligned}
\mathbb{E}[\bar{\sigma}_t^2] \leq & \frac{1}{K} \sum\nolimits_{k=1}^{K} \frac{1}{Q^{\mathrm{M}}} \mathbb{E}[\|\Delta\hat{\mathbf{w}}_{t,k}\|^2] \\
\leq & \frac{1}{L^2 K Q^{\mathrm{M}}} \sum\nolimits_{k=1}^{K} \mathbb{E}\left[\left\|\sum\nolimits_{i=1}^{I_t} \nabla \hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,i-1})\right\|^2\right] \\
\leq & \frac{I_t}{L^2 K Q^{\mathrm{M}}} \sum\nolimits_{k=1}^{K} \sum\nolimits_{i=1}^{I_t} \mathbb{E}[\|\nabla \hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,i-1})\|^2] \\
\leq & \frac{I_t^2 G^2}{L^2 Q^{\mathrm{M}}}.
\end{aligned} \tag{A.7}
$$

By substituting (A.5) and (A.7) into (A.4), while taking the expectation on both sides, we have

$$
\begin{aligned}
& \mathbb{E}[F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t)] \\
\leq & \frac{1}{2L}\left[\hat{\rho}_t \left(\frac{L}{\mu}\gamma^{\mathrm{A}} - I_t\right) - \tilde{\rho}_t I_t\right] \mathbb{E}[\|\nabla F(\mathbf{w}_t)\|^2] \\
& + \hat{\rho}_t \frac{G^2}{L}\left(\frac{1}{2} + \gamma^{\mathrm{A}}\right) I_t^2 + \frac{G^2}{L} I_t^2 + \frac{3\delta^2 + 5G^2}{2L} I_t.
\end{aligned} \tag{A.8}
$$

Based on (28), we have the following PL inequality for $F(\mathbf{w}_t)$:

$$
\mathbb{E}[\|\nabla F(\mathbf{w}_t)\|^2] \geq 2\mu \mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}^*)]. \tag{A.9}
$$

Plugging (A.9) into (A.8) while subtracting $F(\mathbf{w}^*)$ from both sides, we obtain

$$
\begin{aligned}
& \mathbb{E}[F(\mathbf{w}_{t+1}) - F(\mathbf{w}^*)] \\
\leq & (1 - \frac{\mu}{L} I_t + \hat{\rho}_t \gamma^{\mathrm{A}}) \mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}^*)] \\
& + \frac{G^2}{L}\left[\hat{\rho}_t \left(\frac{1}{2} + \gamma^{\mathrm{A}}\right) + 1\right] I_t^2 + \frac{3\delta^2 + 5G^2}{2L} I_t.
\end{aligned} \tag{A.10}
$$

As discussed in Remark 1, the decay rate $\Lambda_{t,1}$ should be limited in the range of $[0,1]$, which ensures $(1/2L)\left[\hat{\rho}_t\left(L\gamma^{\mathrm{A}}/\mu - I_t\right) - \tilde{\rho}_t I_t\right] \leq 0$. Hence, the correctness of applying the PL inequality is guaranteed. Applying $I_t \approx (L/\mu)\log(1/\varepsilon_t)$ to (A.10), we have

$$
\mathbb{E}[F(\mathbf{w}_{t+1}) - F(\mathbf{w}^*)] \leq \Lambda_{t,1} \mathbb{E}[F(\mathbf{w}_t) - F(\mathbf{w}^*)] + \Lambda_{t,2}. \tag{A.11}
$$

Finally, recursively applying the inequality above for t times and then letting $t = T$, we have (34). The proof is complete.

References

[1] J. Zheng, W. Ni, H. Tian, D. Gündüz, T. Q. S. Quek, and Z. Han, "Semi-federated learning: Convergence analysis and optimization of a hybrid learning framework," *IEEE Trans. Wireless Commun.*, 2023, early access, doi: 10.1109/TWC.2023.3270908.