Supplementary Material for the Paper: "Convergence Analysis and Latency Minimization for Retransmission-Based Semi-Federated Learning"

Jingheng Zheng*, Wanli Ni*, Hui Tian*, Wenchao Jiang†, and Tony Q. S. Quek†
*State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications, Beijing, China
†Information System Technology and Design Pillar,
Singapore University of Technology and Design, Singapore
Email: {zhengjh, charleswall, tianhui}@bupt.edu.cn; {wenchao_jiang, tonyquek}@sutd.edu.sg

In the document, we present the detailed derivations for Lemma 1, Lemma 2, and Theorem 1.

Appendix A
Proof of Lemma 1

Define $X_{t,k,n^{\mathrm{D}}} = |\mathbf{b}_t^{\mathrm{H}}\mathbf{h}_{t,k,n^{\mathrm{D}}}^{\mathrm{D}} - 1|^2, \forall k \in \mathcal{K}$. Since $\mathbf{h}_{t,k,n^{\mathrm{D}}}^{\mathrm{D}} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$, the expectations and variances of independent but non-identically distributed random variables $\{X_{t,k,n^{\mathrm{D}}}\}$ are given by

$$\begin{aligned}\mathbb{E}[X_{t,k,n^{\mathrm{D}}}] &= \mathbb{E}[|\mathbf{b}_t^{\mathrm{H}}\mathbf{h}_{t,k,n^{\mathrm{D}}}^{\mathrm{D}}|^2] + 1 \\ &= \|\mathbf{b}_t\|^2 + 1 \\ &\overset{(a)}{=} 2, \forall k \in \mathcal{K},\end{aligned} \tag{42}$$

$$\begin{aligned}\mathbb{D}[X_{t,k,n^{\mathrm{D}}}] &= \mathbb{E}[(|\mathbf{b}_t^{\mathrm{H}}\mathbf{h}_{t,k,n^{\mathrm{D}}}^{\mathrm{D}}|^2 - \|\mathbf{b}_t\|^2 - 2\mathrm{Re}\{\mathbf{b}_t^{\mathrm{H}}\mathbf{h}_{t,k,n^{\mathrm{D}}}^{\mathrm{D}}\})^2] \\ &= \|\mathbf{b}_t\|^2(\|\mathbf{b}_t\|^2 + 2) \\ &\overset{(b)}{=} 3, \forall k \in \mathcal{K},\end{aligned} \tag{43}$$

where (a) and (b) are because $\|\mathbf{b}_t\| = 1$. By defining $\bar{s}_{t,n^{\mathrm{D}}}^2 = \sum_{k=1}^K \mathbb{D}[X_{t,k,n^{\mathrm{D}}}] = 3K$, we have $(\sum_{k=1}^K (X_{t,k,n^{\mathrm{D}}} - \mathbb{E}[X_{t,k,n^{\mathrm{D}}}]))/\bar{s}_{t,n^{\mathrm{D}}} \sim \mathcal{N}(0,1)$ according to Lyapunov's central limit theorem [1]. As a result, it can be obtained that

$$\begin{aligned}&Pr\{\mathrm{MSE}_t^{\mathrm{D}} \le \gamma^{\mathrm{D}}\} \\ =&Pr\left\{\frac{\sum_{k=1}^K (X_{t,k,n^{\mathrm{D}}} - \mathbb{E}[X_{t,k,n^{\mathrm{D}}}])}{\bar{s}_{t,n^{\mathrm{D}}}} \le \frac{K^2\gamma^{\mathrm{D}} - 2K - \frac{\tilde{\sigma}^2}{\zeta_t}}{\sqrt{3K}}\right\} \\ \approx&\Phi\left(\frac{1}{\sqrt{3K}}\left(K^2\gamma^{\mathrm{D}} - 2K - \frac{\tilde{\sigma}^2}{\zeta_t}\right)\right).\end{aligned} \tag{44}$$

The proof is complete.

Appendix B
Proof of Lemma 2

For two consecutive FL iterations of the k-th device in the t-th round, by substituting (2) into (23), we have

$$\begin{aligned}&\hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,i}) - \hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,i-1}) \\ &\le \left(\frac{L}{2}\hat{\eta}_t^2 - \hat{\eta}_t\right)\|\nabla\hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,i-1})\|^2 \\ &\overset{(a)}{=} -\frac{1}{2L}\|\nabla\hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,i-1})\|^2,\end{aligned} \tag{45}$$

where (a) comes from setting $\hat{\eta}_t = 1/L$. Based on Assumption 2, one can derive the celebrated PL inequality, given by [2]

$$\begin{aligned}&\|\nabla\hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,i-1})\|^2 \\ &\ge 2\mu[\hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,i-1}) - \hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k}^*)],\end{aligned} \tag{46}$$

where $\hat{\mathbf{w}}_{t,k}^*$ denotes the optimal model regarding the loss function $\hat{F}_{t,k}(\mathbf{w})$. After applying (46) to (45), while subtracting $\hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k}^*)$ from both sides of the result, we have

$$\begin{aligned}&\hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,i}) - \hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k}^*) \\ &\le \left(1 - \frac{\mu}{L}\right)[\hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,i-1}) - \hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k}^*)].\end{aligned} \tag{47}$$

Recursively applying (47) for i times, while taking the expectation of both sides, it is obtained that

$$\begin{aligned}&\mathbb{E}[\hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,i}) - \hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k}^*)] \\ &\overset{(b)}{\le} \left(1 - \frac{\mu}{L}\right)^i \mathbb{E}[\hat{F}_{t,k}(\mathbf{w}_t) - \hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k}^*)],\end{aligned} \tag{48}$$

where (b) is because $\hat{\mathbf{w}}_{t,k,0} = \mathbf{w}_t$. Given a local target accuracy $\hat{\varepsilon}_t$, the convergence requirement is mathematically described as

$$\begin{aligned}&\mathbb{E}[\hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,i}) - \hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k}^*)] \\ &\le \hat{\varepsilon}_t\mathbb{E}[\hat{F}_{t,k}(\mathbf{w}_t) - \hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k}^*)].\end{aligned} \tag{49}$$

Based on (48), meeting the requirement in (49) is equivalent to ensuring the condition $(1 - \mu/L)^i \le \hat{\varepsilon}_t$. Considering the inequality $1 - x \le e^{-x}, \forall x \in \mathbb{R}$, the aforementioned condition holds if $e^{-(\mu/L)} \le \hat{\varepsilon}_t$, i.e.,

$$i \ge \frac{L}{\mu}\log\left(\frac{1}{\hat{\varepsilon}_t}\right). \tag{50}$$

Replacing i with $\hat{I}_t$, (29) can be obtained.

As for CL, after plugging (5) into (23), while taking the expectation on both sides of the resultant inequality, we have

$$\begin{aligned}
&\mathbb{E}[\tilde{F}_t(\tilde{\mathbf{w}}_{t,i}) - \tilde{F}_t(\tilde{\mathbf{w}}_{t,i-1})] \\
&\leq -\tilde{\eta}_t \nabla \tilde{F}_t(\tilde{\mathbf{w}}_{t,i-1})^{\mathrm{T}} \mathbb{E}[\nabla \bar{F}_t(\tilde{\mathbf{w}}_{t,i-1})] \\
&\quad + \frac{L}{2}\tilde{\eta}_t^2 \mathbb{E}[\|\nabla \bar{F}_t(\tilde{\mathbf{w}}_{t,i-1})\|^2] \\
&\overset{(c)}{\leq} -\tilde{\eta}_t \mathbb{E}[\|\nabla \tilde{F}_t(\tilde{\mathbf{w}}_{t,i-1})\|^2] \\
&\quad + \frac{L}{2}\tilde{\eta}_t^2 \mathbb{E}[\|\nabla \bar{F}_t(\tilde{\mathbf{w}}_{t,i-1})\|^2], \qquad (51)
\end{aligned}$$

where (c) is because $\nabla \bar{F}_t(\mathbf{w})$ provides an unbiased estimation of $\nabla \tilde{F}_t(\mathbf{w})$, as presented in Assumption 3. Besides, it is noticed that

$$\begin{aligned}
&\mathbb{E}[\|\nabla \bar{F}_t(\tilde{\mathbf{w}}_{t,i-1}) - \nabla \bar{F}_t(\tilde{\mathbf{w}}_t^*)\|^2] \\
=&\mathbb{E}[\|\nabla \bar{F}_t(\tilde{\mathbf{w}}_{t,i-1})\|^2] + \mathbb{E}[\|\nabla \bar{F}_t(\tilde{\mathbf{w}}_t^*)\|^2] \\
&- 2\mathbb{E}[\nabla \bar{F}_t(\tilde{\mathbf{w}}_t^*)]^{\mathrm{T}} \mathbb{E}[\nabla \bar{F}_t(\tilde{\mathbf{w}}_{t,i-1})] \\
\overset{(d)}{=}&\mathbb{E}[\|\nabla \bar{F}_t(\tilde{\mathbf{w}}_{t,i-1})\|^2] + \mathbb{E}[\|\nabla \bar{F}_t(\tilde{\mathbf{w}}_t^*)\|^2], \qquad (52)
\end{aligned}$$

where $\tilde{\mathbf{w}}_t^*$ denotes the optimal model of the loss function $\tilde{F}_t(\mathbf{w})$ and (d) is because $\mathbb{E}[\nabla \bar{F}_t(\tilde{\mathbf{w}}_t^*)] = \nabla \tilde{F}_t(\tilde{\mathbf{w}}_t^*) = \mathbf{0}$. Since $\mathbb{E}[\|\nabla \bar{F}_t(\tilde{\mathbf{w}}_t^*)\|^2] \geq 0$, we derive the following inequality from (52):

$$\begin{aligned}
&\mathbb{E}[\|\nabla \bar{F}_t(\tilde{\mathbf{w}}_{t,i-1})\|^2] \\
\leq&\mathbb{E}[\|\nabla \bar{F}_t(\tilde{\mathbf{w}}_{t,i-1}) - \nabla \bar{F}_t(\tilde{\mathbf{w}}_t^*)\|^2] \\
\overset{(e)}{\leq}&L^2 \mathbb{E}[\|\tilde{\mathbf{w}}_{t,i-1} - \tilde{\mathbf{w}}_t^*\|^2] \\
\overset{(f)}{\leq}&\frac{\tilde{D}_t^{\mathrm{BS}} L^2}{\bar{D}_t} \mathbb{E}[\|\tilde{\mathbf{w}}_{t,i-1} - \tilde{\mathbf{w}}_t^*\|^2], \qquad (53)
\end{aligned}$$

where (e) stems from Assumption 1 and (f) comes from $\tilde{D}_t^{\mathrm{BS}}/\bar{D}_t \geq 1$. Based on Assumption 2, one can have

$$\|\tilde{\mathbf{w}}_{t,i-1} - \tilde{\mathbf{w}}_t^*\|^2 \leq \frac{2}{\mu}[\tilde{F}_t(\tilde{\mathbf{w}}_{t,i-1}) - \tilde{F}_t(\tilde{\mathbf{w}}_t^*)]. \qquad (54)$$

By plugging (54) into (53), we bound $\mathbb{E}[\|\nabla \bar{F}_t(\tilde{\mathbf{w}}_{t,i-1})\|^2]$ by

$$\mathbb{E}[\|\nabla \bar{F}_t(\tilde{\mathbf{w}}_{t,i-1})\|^2] \leq \frac{2\tilde{D}_t^{\mathrm{BS}} L^2}{\bar{D}_t \mu} \mathbb{E}[\tilde{F}_t(\tilde{\mathbf{w}}_{t,i-1}) - \tilde{F}_t(\tilde{\mathbf{w}}_t^*)]. \qquad (55)$$

Applying (55) and the PL inequality in (46) to (51), we have

$$\begin{aligned}
&\mathbb{E}[\tilde{F}_t(\tilde{\mathbf{w}}_{t,i}) - \tilde{F}_t(\tilde{\mathbf{w}}_{t,i-1})] \\
&\leq \left(\frac{\tilde{D}_t^{\mathrm{BS}} L^3}{\bar{D}_t \mu}\tilde{\eta}_t^2 - 2\mu\tilde{\eta}_t\right) \mathbb{E}[\tilde{F}_t(\tilde{\mathbf{w}}_{t,i-1}) - \tilde{F}_t(\tilde{\mathbf{w}}_t^*)] \\
&\overset{(g)}{\leq} -\frac{\bar{D}_t \mu^3}{\tilde{D}_t^{\mathrm{BS}} L^3} \mathbb{E}[\tilde{F}_t(\tilde{\mathbf{w}}_{t,i-1}) - \tilde{F}_t(\tilde{\mathbf{w}}_t^*)], \qquad (56)
\end{aligned}$$

where (g) is achieved by setting $\tilde{\eta}_t = (\bar{D}_t\mu^2)/(\tilde{D}_t^{\mathrm{BS}} L^3)$. By subtracting $\tilde{F}_t(\tilde{\mathbf{w}}_t^*)$ from both sides of (56) and recursively applying the result for i times, one can find that

$$\begin{aligned}
&\mathbb{E}[\tilde{F}_t(\tilde{\mathbf{w}}_{t,i}) - \tilde{F}_t(\tilde{\mathbf{w}}_t^*)] \\
&\leq \left(1 - \frac{\bar{D}_t \mu^3}{\tilde{D}_t^{\mathrm{BS}} L^3}\right)^i \mathbb{E}[\tilde{F}_t(\mathbf{w}_t) - \tilde{F}_t(\tilde{\mathbf{w}}_t^*)]. \qquad (57)
\end{aligned}$$

Given a local target accuracy $\tilde{\varepsilon}_t$, while applying $1 - x \leq e^x$, one can ensure

$$\mathbb{E}[\tilde{F}_t(\tilde{\mathbf{w}}_{t,i}) - \tilde{F}_t(\tilde{\mathbf{w}}_t^*)] \leq \tilde{\varepsilon}_t \mathbb{E}[\tilde{F}_t(\mathbf{w}_t) - \tilde{F}_t(\tilde{\mathbf{w}}_t^*)], \qquad (58)$$

if the following condition is met:

$$i \geq \frac{\tilde{D}_t^{\mathrm{BS}} L^3}{\bar{D}_t \mu^3} \log\left(\frac{1}{\tilde{\varepsilon}_t}\right). \qquad (59)$$

Substituting i with $\tilde{I}_t$, we reach (30). The proof is complete.

Appendix C
Proof of Theorem 1

Based on Assumption 1, by plugging $\mathbf{w} = \mathbf{w}_{t+1}$ and $\mathbf{w}' = \mathbf{w}_t$ into (23), we have

$$F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t) \leq (\mathbf{w}_{t+1} - \mathbf{w}_t)^{\mathrm{T}} \nabla F(\mathbf{w}_t) + \frac{L}{2}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2. \qquad (60)$$

Based on (7), one can derive that

$$\begin{aligned}
\mathbf{w}_{t+1} - \mathbf{w}_t =& \hat{\rho}_t \Delta\hat{\mathbf{w}}_t + \tilde{\rho}_t \Delta\tilde{\mathbf{w}}_t \\
=& -\frac{\hat{\rho}_t\hat{\eta}_t}{K}\sum_{k=1}^{K}\sum_{i=1}^{\hat{I}_t} \nabla \hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,i-1}) \\
& - \tilde{\rho}_t\tilde{\eta}_t \sum_{i=1}^{\tilde{I}_t} \nabla \bar{F}_t(\tilde{\mathbf{w}}_{t,i-1}). \qquad (61)
\end{aligned}$$

Plugging (61) into the first term on the right-hand side of (60), while taking the expectation on both sides, we have

$$\begin{aligned}
&\mathbb{E}[F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t)] \\
\overset{(a)}{\leq}& -\frac{\hat{\rho}_t\hat{\eta}_t}{K}\sum_{k=1}^{K}\sum_{i=1}^{\hat{I}_t} \nabla F(\mathbf{w}_t)^{\mathrm{T}} \nabla \hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,i-1}) \\
& -\tilde{\rho}_t\tilde{\eta}_t \sum_{i=1}^{\tilde{I}_t} \nabla F(\mathbf{w}_t)^{\mathrm{T}} \nabla \tilde{F}_t(\tilde{\mathbf{w}}_{t,i-1}) + \frac{L}{2}\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2] \\
\overset{(b)}{=}& -\frac{\hat{\rho}_t\hat{\eta}_t\hat{I}_t + \tilde{\rho}_t\tilde{\eta}_t\tilde{I}_t}{2}\|\nabla F(\mathbf{w}_t)\|^2 \\
& - \frac{\hat{\rho}_t\hat{\eta}_t}{2K}\sum_{k=1}^{K}\sum_{i=1}^{\hat{I}_t} \|\nabla \hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,i-1}) - \nabla \hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k}^*)\|^2 \\
& + \frac{\hat{\rho}_t\hat{\eta}_t}{2K}\sum_{k=1}^{K}\sum_{i=1}^{\hat{I}_t} \|\nabla F(\mathbf{w}_t) - \nabla \hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,i-1})\|^2 \\
& - \frac{\tilde{\rho}_t\tilde{\eta}_t}{2}\sum_{i=1}^{\tilde{I}_t} \|\nabla \tilde{F}_t(\tilde{\mathbf{w}}_{t,i-1}) - \nabla \tilde{F}_t(\tilde{\mathbf{w}}_t^*)\|^2 \\
& + \frac{\tilde{\rho}_t\tilde{\eta}_t}{2}\sum_{i=1}^{\tilde{I}_t} \|\nabla F(\mathbf{w}_t) - \nabla \tilde{F}_t(\tilde{\mathbf{w}}_{t,i-1})\|^2 \\
& + \frac{L}{2}\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2] \\
\overset{(c)}{\leq}& -\frac{\hat{\rho}_t\hat{\eta}_t\hat{I}_t + \tilde{\rho}_t\tilde{\eta}_t\tilde{I}_t}{2}\|\nabla F(\mathbf{w}_t)\|^2 \\
& - \frac{\hat{\rho}_t\hat{\eta}_t}{2K}\sum_{k=1}^{K} \|\nabla \hat{F}_{t,k}(\mathbf{w}_t) - \nabla \hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k}^*)\|^2 \\
& + \frac{\hat{\rho}_t\hat{\eta}_t}{2K}\sum_{k=1}^{K}\sum_{i=1}^{\hat{I}_t} \|\nabla F(\mathbf{w}_t) - \nabla \hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,i-1})\|^2 \\
& - \frac{\tilde{\rho}_t\tilde{\eta}_t}{2}\|\nabla \tilde{F}_t(\mathbf{w}_t) - \nabla \tilde{F}_t(\tilde{\mathbf{w}}_t^*)\|^2 \\
& + \frac{\tilde{\rho}_t\tilde{\eta}_t}{2}\sum_{i=1}^{\tilde{I}_t} \|\nabla F(\mathbf{w}_t) - \nabla \tilde{F}_t(\tilde{\mathbf{w}}_{t,i-1})\|^2 \\
& + \frac{L}{2}\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2], \qquad (62)
\end{aligned}$$

where (a) comes from applying the unbiased estimation in Assumption 3, (b) is because $-\nabla F(\mathbf{w}_t)^{\mathrm{T}} \nabla \hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,i-1}) =$

$-(1/2)(\|\nabla F(\mathbf{w}_t)\|^2 + \|\nabla \hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,i-1})\|^2 - \|\nabla F(\mathbf{w}_t) - \nabla \hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,i-1})\|^2)$ and $-\nabla F(\mathbf{w}_t)^{\mathrm{T}} \nabla \tilde{F}_t(\tilde{\mathbf{w}}_{t,i-1}) = -(1/2)(\|\nabla F(\mathbf{w}_t)\|^2 + \|\nabla \tilde{F}_t(\tilde{\mathbf{w}}_{t,i-1})\|^2 - \|\nabla F(\mathbf{w}_t) - \nabla \tilde{F}_t(\tilde{\mathbf{w}}_{t,i-1})\|^2)$, and (c) stems from removing $-(\hat{\rho}_t\hat{\eta}_t)/(2K)\sum_{k=1}^{K}\sum_{i=2}^{\hat{I}_t}\|\nabla \hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,i-1}) - \nabla \hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k}^*)\|^2$ and $-(\tilde{\rho}_t\tilde{\eta}_t/2)\sum_{i=2}^{\tilde{I}_t}\|\nabla \tilde{F}_t(\tilde{\mathbf{w}}_{t,i-1}) - \nabla \tilde{F}_t(\tilde{\mathbf{w}}_t^*)\|^2$ from the right-hand side.

Based on Assumption 2, we have

$$\|\nabla \hat{F}_{t,k}(\mathbf{w}_t) - \nabla \hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k}^*)\|^2 \geq \mu^2\|\mathbf{w}_t - \hat{\mathbf{w}}_{t,k}^*\|^2 \tag{63}$$
$$\|\nabla \tilde{F}_t(\mathbf{w}_t) - \nabla \tilde{F}_t(\tilde{\mathbf{w}}_t^*)\|^2 \geq \mu^2\|\mathbf{w}_t - \tilde{\mathbf{w}}_t^*\|^2. \tag{64}$$

Applying (63) and (64) to (62), we have

$$\begin{aligned}
&\mathbb{E}[F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t)]\\
\overset{(d)}{\leq} & -\frac{\hat{\rho}_t\hat{\eta}_t\hat{I}_t + \tilde{\rho}_t\tilde{\eta}_t\tilde{I}_t}{2}\|\nabla F(\mathbf{w}_t)\|^2 + \frac{L}{2}\mathbb{E}[\|\mathbf{w}_{t+1}-\mathbf{w}_t\|^2]\\
& -\frac{\hat{\rho}_t\hat{\eta}_t\mu^2}{2K}\sum_{k=1}^{K}\|\mathbf{w}_t - \hat{\mathbf{w}}_{t,k}^*\|^2 - \frac{\tilde{\rho}_t\tilde{\eta}_t\mu^2}{2}\|\mathbf{w}_t - \tilde{\mathbf{w}}_t^*\|^2\\
& + \frac{\hat{\rho}_t\hat{\eta}_t}{K}\sum_{k=1}^{K}\sum_{i=1}^{\hat{I}_t}\Big(\|\nabla F(\mathbf{w}_t) - \nabla \hat{F}_{t,k}(\mathbf{w}_t)\|^2\\
& + \|\nabla \hat{F}_{t,k}(\mathbf{w}_t) - \nabla \hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,i-1})\|^2\Big)\\
& + \tilde{\rho}_t\tilde{\eta}_t\sum_{i=1}^{\tilde{I}_t}\Big(\|\nabla F(\mathbf{w}_t) - \nabla \tilde{F}_t(\mathbf{w}_t)\|^2\\
& + \|\nabla \tilde{F}_t(\mathbf{w}_t) - \nabla \tilde{F}_t(\tilde{\mathbf{w}}_{t,i-1})\|^2\Big)\\
\overset{(e)}{\leq} & -\frac{\hat{\rho}_t\hat{\eta}_t\hat{I}_t + \tilde{\rho}_t\tilde{\eta}_t\tilde{I}_t}{2}\|\nabla F(\mathbf{w}_t)\|^2 + \frac{L}{2}\mathbb{E}[\|\mathbf{w}_{t+1}-\mathbf{w}_t\|^2]\\
& -\frac{\hat{\rho}_t\hat{\eta}_t\mu^2}{2K}\sum_{k=1}^{K}\|\mathbf{w}_t - \hat{\mathbf{w}}_{t,k}^*\|^2 - \frac{\tilde{\rho}_t\tilde{\eta}_t\mu^2}{2}\|\mathbf{w}_t - \tilde{\mathbf{w}}_t^*\|^2\\
& + \hat{\rho}_t\hat{\eta}_t\hat{I}_t\hat{\delta}^2 + \tilde{\rho}_t\tilde{\eta}_t\tilde{I}_t\tilde{\delta}^2\\
& + \frac{\hat{\rho}_t\hat{\eta}_t}{K}\sum_{k=1}^{K}\sum_{i=1}^{\hat{I}_t}\|\nabla \hat{F}_{t,k}(\mathbf{w}_t) - \nabla \hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,i-1})\|^2\\
& + \tilde{\rho}_t\tilde{\eta}_t\sum_{i=1}^{\tilde{I}_t}\|\nabla \tilde{F}_t(\mathbf{w}_t) - \nabla \tilde{F}_t(\tilde{\mathbf{w}}_{t,i-1})\|^2,
\end{aligned} \tag{65}$$

where (d) is because of the Cauchy-Schwarz inequality and (e) applies the inequalities in Assumption 4.

Based on Assumption 1, we now bound $\|\nabla \hat{F}_{t,k}(\mathbf{w}_t) - \nabla \hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,i-1})\|^2$ and $\|\nabla \tilde{F}_t(\mathbf{w}_t) - \nabla \tilde{F}_t(\tilde{\mathbf{w}}_{t,i-1})\|^2$ by

$$\begin{aligned}
&\|\nabla \hat{F}_{t,k}(\mathbf{w}_t) - \nabla \hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,i-1})\|^2\\
\leq & L^2\|\mathbf{w}_t - \hat{\mathbf{w}}_{t,k,i-1}\|^2\\
= & L^2\hat{\eta}_t^2\Big\|\sum_{j=1}^{i-1}\nabla \hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,j-1})\Big\|^2\\
\overset{(f)}{\leq} & L^2\hat{\eta}_t^2(i-1)\sum_{j=1}^{i-1}\|\nabla \hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,j-1})\|^2\\
\overset{(g)}{\leq} & L^2\hat{\eta}_t^2(i-1)^2\hat{G}^2,
\end{aligned} \tag{66}$$

$$\begin{aligned}
&\|\nabla \tilde{F}_t(\mathbf{w}_t) - \nabla \tilde{F}_t(\tilde{\mathbf{w}}_{t,i-1})\|^2\\
\leq & L^2\|\mathbf{w}_t - \tilde{\mathbf{w}}_{t,i-1}\|^2\\
= & L^2\tilde{\eta}_t^2\Big\|\sum_{j=1}^{i-1}\nabla \tilde{F}_t(\tilde{\mathbf{w}}_{t,j-1})\Big\|^2\\
\leq & L^2\tilde{\eta}_t^2(i-1)^2\tilde{G}^2,
\end{aligned} \tag{67}$$

where (f) comes from using the Cauchy-Schwarz inequality and (g) is because of Assumption 5. Since $\sum_{j=1}^{i-1} j^2 = i(i-1)(2i-1)/6$, by substituting (66) and (67) into (65), we have

$$\begin{aligned}
&\mathbb{E}[F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t)]\\
\leq & -\frac{\hat{\rho}_t\hat{\eta}_t\hat{I}_t + \tilde{\rho}_t\tilde{\eta}_t\tilde{I}_t}{2}\|\nabla F(\mathbf{w}_t)\|^2\\
& -\frac{\hat{\rho}_t\hat{\eta}_t\mu^2}{2K}\sum_{k=1}^{K}\|\mathbf{w}_t - \hat{\mathbf{w}}_{t,k}^*\|^2 - \frac{\tilde{\rho}_t\tilde{\eta}_t\mu^2}{2}\|\mathbf{w}_t - \tilde{\mathbf{w}}_t^*\|^2\\
& + \hat{\rho}_t\hat{\eta}_t\hat{I}_t\left[\hat{\delta}^2 + L^2\hat{\eta}_t^2\hat{G}^2\frac{(\hat{I}_t-1)(2\hat{I}_t-1)}{6}\right]\\
& + \tilde{\rho}_t\tilde{\eta}_t\tilde{I}_t\left[\tilde{\delta}^2 + L^2\tilde{\eta}_t\tilde{G}^2\frac{(\tilde{I}_t-1)(2\tilde{I}_t-1)}{6}\right]\\
& + \frac{L}{2}\mathbb{E}[\|\mathbf{w}_{t+1}-\mathbf{w}_t\|^2].
\end{aligned} \tag{68}$$

Then, we expand and bound $\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2$, given by

$$\begin{aligned}
&\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2\\
= & \Big\|\frac{\hat{\rho}_t}{K}\sum_{k=1}^{K}(\hat{\mathbf{w}}_{t,k,\hat{I}_t} - \hat{\mathbf{w}}_{t,k}^*) + \tilde{\rho}_t(\tilde{\mathbf{w}}_{t,\tilde{I}} - \tilde{\mathbf{w}}_t^*)\\
& + \frac{\hat{\rho}_t}{K}\sum_{k=1}^{K}(\hat{\mathbf{w}}_{t,k}^* - \mathbf{w}_t) + \tilde{\rho}_t(\tilde{\mathbf{w}}_t^* - \mathbf{w}_t)\Big\|^2\\
\overset{(h)}{\leq} & 2\left(\frac{\hat{\rho}^2}{K} + \tilde{\rho}_t^2\right)\Big(\sum_{k=1}^{K}\|\hat{\mathbf{w}}_{t,k,\hat{I}_t} - \hat{\mathbf{w}}_{t,k}^*\|^2 + \|\tilde{\mathbf{w}}_{t,\tilde{I}} - \tilde{\mathbf{w}}_t^*\|^2\\
& + \sum_{k=1}^{K}\|\mathbf{w}_t - \hat{\mathbf{w}}_{t,k}^*\|^2 + \|\mathbf{w}_t - \tilde{\mathbf{w}}_t^*\|^2\Big),
\end{aligned} \tag{69}$$

where (h) is because of the triangle inequality and the Cauchy-Schwarz inequality. Plugging (69) into (68), while setting $\hat{\eta}_t = 1/L$ and $\tilde{\eta}_t = \bar{D}_t\mu^2/(\tilde{D}_t^{\mathrm{BS}}L^3)$, it is obtained that

$$\begin{aligned}
&\mathbb{E}[F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t)]\\
\leq & -\left(\frac{\mu\hat{\rho}_t\hat{I}_t}{L} + \frac{\mu^3\bar{D}_t\tilde{\rho}_t\tilde{I}_t}{L^3\tilde{D}_t^{\mathrm{BS}}}\right)\|\nabla F(\mathbf{w}_t)\|^2 + \hat{\xi}_t + \tilde{\xi}_t\\
& + L\left(\frac{\hat{\rho}_t^2}{K} + \tilde{\rho}_t^2\right)\Big(\sum_{k=1}^{K}\|\hat{\mathbf{w}}_{t,k,\hat{I}_t} - \hat{\mathbf{w}}_{t,k}^*\|^2 + \|\tilde{\mathbf{w}}_{t,\tilde{I}_t} - \tilde{\mathbf{w}}_t^*\|^2\Big)\\
& + \left[L\left(\frac{\hat{\rho}_t^2}{K} + \tilde{\rho}_t^2\right) - \frac{\hat{\rho}_t\hat{\eta}_t\mu^2}{2K}\right]\sum_{k=1}^{K}\|\mathbf{w}_t - \hat{\mathbf{w}}_{t,k}^*\|^2\\
& + \left[L\left(\frac{\hat{\rho}_t^2}{K} + \tilde{\rho}_t^2\right) - \frac{\tilde{\rho}_t\tilde{\eta}_t\mu^2}{2}\right]\|\mathbf{w}_t - \tilde{\mathbf{w}}_t^*\|^2,
\end{aligned} \tag{70}$$

where the definitions of $\hat{\xi}_t$ and $\tilde{\xi}_t$ are presented in Theorem 1.

Based on Assumption 2, we have

$$\|\mathbf{w}_t - \hat{\mathbf{w}}_{t,k}^*\|^2 \leq \frac{2}{\mu}[\hat{F}_{t,k}(\mathbf{w}_t) - \hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k}^*)], \tag{71}$$
$$\|\mathbf{w}_t - \tilde{\mathbf{w}}_t^*\|^2 \leq \frac{2}{\mu}[\tilde{F}_t(\mathbf{w}_t) - \tilde{F}_t(\tilde{\mathbf{w}}_t^*)]. \tag{72}$$

Again, by invoking Lemma 2, one can have the following inequalities based on Assumption 2:

$$\begin{aligned}
\|\hat{\mathbf{w}}_{t,k,\hat{I}_t} - \hat{\mathbf{w}}_{t,k}^*\|^2 \leq & \frac{2}{\mu}[\hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k,\hat{I}_t}) - \hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k}^*)]\\
\leq & \frac{2\hat{\varepsilon}_t}{\mu}[\hat{F}_{t,k}(\mathbf{w}_t) - \hat{F}_{t,k}(\hat{\mathbf{w}}_{t,k}^*)],
\end{aligned} \tag{73}$$

$$\begin{aligned}\|\tilde{\mathbf{w}}_{t,\tilde{I}_t}-\tilde{\mathbf{w}}_t^*\|^2 \leq & \frac{2}{\mu}[\tilde{F}_t(\tilde{\mathbf{w}}_{t,\tilde{I}_t})-\tilde{F}_t(\tilde{\mathbf{w}}_t^*)] \\ \leq & \frac{2\tilde{\varepsilon}_t}{\mu}[\tilde{F}_t(\mathbf{w}_t)-\tilde{F}_t(\tilde{\mathbf{w}}_t^*)]. \end{aligned} \tag{74}$$

Plugging (71)–(74) into (70), while applying the notations defined in Theorem 1, we have

$$\begin{aligned}&\mathbb{E}[F(\mathbf{w}_{t+1})-F(\mathbf{w}_t)] \\ \leq & -\left(\frac{\mu\hat{\rho}_t\hat{I}_t}{L}+\frac{\mu^3\bar{D}_t\tilde{\rho}_t\tilde{I}_t}{L^3\tilde{D}_t^{\mathrm{BS}}}\right)\|\nabla F(\mathbf{w}_t)\|^2+\hat{\xi}_t+\tilde{\xi}_t \\ & +\hat{\phi}_t\sum\nolimits_{k=1}^{K}\Delta\hat{F}_{t,k}(\mathbf{w}_t)+\tilde{\phi}_t\Delta\tilde{F}_t(\mathbf{w}_t). \end{aligned} \tag{75}$$

Based on (46), after applying the PL inequality regarding $F(\mathbf{w})$ to (75), we have

$$\mathbb{E}[F(\mathbf{w}_{t+1})-F(\mathbf{w}^*)] \leq \Lambda_{1,t}\mathbb{E}[F(\mathbf{w}_t)-F(\mathbf{w}^*)]+\Lambda_{2,t}. \tag{76}$$

Recursively applying (76) for t times and setting $t=T$, we have (31) eventually. The proof is complete.

References

[1] A. Kammoun, M. Kharouf, W. Hachem, and J. Najim, "A central limit theorem for the SINR at the LMMSE estimator output for large-dimensional signals," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 5048–5063, Nov. 2009.

[2] W. Ni, Y. Liu, Y. C. Eldar, Z. Yang, and H. Tian, "STAR-RIS integrated nonorthogonal multiple access and over-the-air federated learning: Framework, analysis, and optimization," *IEEE Internet Things J.*, vol. 9, no. 18, pp. 17 136–17 156, Sep. 2022.