

# Supplementary Material: Semi-Asynchronous Federated Split Learning for Computing-Limited Devices in Wireless Networks

Huiqing Ao, Hui Tian, Wanli Ni, Gaofeng Nie, and Dusit Niyato, *Fellow, IEEE*

## APPENDIX A PROOF OF THEOREM 1

We define  $\mathbf{g}_{n,t} = \sum_{\tau=0}^{H-1} \mathbf{g}_{n,t}^\tau$  and  $\bar{\mathbf{g}}_{n,t} = \sum_{\tau=0}^{H-1} \bar{\mathbf{g}}_{n,t}^\tau$ , thus  $\mathbb{E}[\mathbf{g}_{n,t}] = \bar{\mathbf{g}}_{n,t}$ . We also define the optimal model parameter  $\mathbf{w}^* = \arg \min_{\mathbf{w}} F(\mathbf{w})$ ,  $\mathbf{w}_{n,t} = \mathbf{w}_{n,t}^0$ , and  $\mathbf{w}_{n,t+1} = \mathbf{w}_{n,t}^H$ . Since the global loss function  $F(\mathbf{w})$  is  $\ell$ -smooth, we have

$$\begin{aligned} \mathbb{E}[F(\mathbf{w}_{t+1})] - F^* &\leq \frac{\ell}{2} \mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2] \\ &= \frac{\ell}{2} \mathbb{E}[\|\sum_{n=1}^N m_{n,t} \rho_{n,t} \mathbf{w}_{n,t+1} - \mathbf{w}^*\|^2] \\ &\stackrel{(a)}{=} \frac{\ell}{2} \mathbb{E}[\|\sum_{n=1}^N m_{n,t} \rho_{n,t} (\mathbf{w}_{n,t+1} - \mathbf{w}^*)\|^2] \\ &\stackrel{(b)}{\leq} \frac{\ell N}{2} \sum_{n=1}^N \mathbb{E}[\|m_{n,t} \rho_{n,t} (\mathbf{w}_{n,t+1} - \mathbf{w}^*)\|^2] \\ &= \frac{\ell N}{2} \sum_{n=1}^N m_{n,t}^2 \rho_{n,t}^2 \mathbb{E}[\|\mathbf{w}_{n,t+1} - \mathbf{w}^*\|^2], \end{aligned} \quad (1)$$

where (a) follows the fact that  $\sum_{n=1}^N m_{n,t} \rho_{n,t} = 1$ , and (b) follows by applying the basic inequality  $\|\sum_{n=1}^N \mathbf{a}_n\|^2 \leq N \sum_{n=1}^N \|\mathbf{a}_n\|^2$ . Now, we can bound  $\mathbb{E}[\|\mathbf{w}_{n,t+1} - \mathbf{w}^*\|^2]$  by applying the following lemmas.

**Lemma 1.** *Under Assumptions 1 and 2, for any  $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^d$ , we have  $\langle \nabla F_n(\mathbf{x}), \mathbf{z} - \mathbf{y} \rangle \geq F_n(\mathbf{z}) - F_n(\mathbf{y}) + \frac{\mu}{4} \|\mathbf{y} - \mathbf{z}\|^2 - \ell \|\mathbf{z} - \mathbf{x}\|^2$ ,  $\forall n$ .*

**Lemma 2.** *Under Assumption 3, then the variance of  $\mathbf{g}_{n,t}$  fulfills  $\mathbb{E}\|\mathbf{g}_{n,t} - \bar{\mathbf{g}}_{n,t}\|^2 \leq H^2 \delta^2$ ,  $\forall n, \forall t$ .*

*Proof:* Since the variance of stochastic gradients is bounded, we have

$$\begin{aligned} \mathbb{E}\|\mathbf{g}_{n,t} - \bar{\mathbf{g}}_{n,t}\|^2 &= \mathbb{E}[\|\sum_{\tau=0}^{H-1} \nabla F_n(\mathbf{w}_{n,t}^\tau) - \sum_{\tau=0}^{H-1} \nabla F_n(\mathbf{w}_{n,t}^\tau; \mathcal{B}_{n,t}^\tau)\|^2] \end{aligned}$$

H. Ao, H. Tian, and G. Nie are with State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: hqao@bupt.edu.cn; tianhui@bupt.edu.cn; niegaofeng@bupt.edu.cn).

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

D. Niyato is with College of Computing and Data Science, Nanyang Technological University, Singapore (email: dniyato@ntu.edu.sg).

$$\begin{aligned} &= \mathbb{E}[\|\sum_{\tau=0}^{H-1} (\nabla F_n(\mathbf{w}_{n,t}^\tau) - \nabla F_n(\mathbf{w}_{n,t}^\tau; \mathcal{B}_{n,t}^\tau))\|^2] \\ &\stackrel{(a)}{\leq} H \sum_{\tau=0}^{H-1} \mathbb{E}[\|\nabla F_n(\mathbf{w}_{n,t}^\tau) - \nabla F_n(\mathbf{w}_{n,t}^\tau; \mathcal{B}_{n,t}^\tau)\|^2] \\ &\stackrel{(b)}{\leq} H^2 \delta^2, \end{aligned} \quad (2)$$

where (a) follows by the  $\|\sum_{n=1}^N \mathbf{a}_n\|^2 \leq N \sum_{n=1}^N \|\mathbf{a}_n\|^2$ , and (b) follows Assumption 3.

**Lemma 3.** *Under Lemma 1, Assumptions 3 and 4, we have  $\mathbb{E}\|\mathbf{w}_{n,t} - \mathbf{w}^* - \eta_t \bar{\mathbf{g}}_{n,t}\|^2 \leq (1 - \frac{\mu H \eta_t}{2}) \mathbb{E}\|\mathbf{w}_{n,t} - \mathbf{w}^*\|^2 + 4H\Lambda\eta_t + \frac{H(H-1)(2H-1)}{3} \ell \eta_t^3 G^2 + 2\eta_t^2 H^2 (\delta^2 + G^2)$ ,  $\forall n, \forall t$ .*

*Proof:* The proof is presented in Appendix C.  $\square$

Hence,  $\mathbb{E}[\|\mathbf{w}_{n,t+1} - \mathbf{w}^*\|^2]$  with Lemmas 2 and 3 can be rewritten as

$$\begin{aligned} &\mathbb{E}[\|\mathbf{w}_{n,t+1} - \mathbf{w}^*\|^2] \\ &= \mathbb{E}[\|\mathbf{w}_{n,t} - \eta_t \mathbf{g}_{n,t} - \mathbf{w}^*\|^2] \\ &= \mathbb{E}[\|\mathbf{w}_{n,t} - \mathbf{w}^* - \eta_t \mathbf{g}_{n,t} - \eta_t \bar{\mathbf{g}}_{n,t} + \eta_t \bar{\mathbf{g}}_{n,t}\|^2] \\ &= \mathbb{E}[\|\mathbf{w}_{n,t} - \mathbf{w}^* - \eta_t \bar{\mathbf{g}}_{n,t}\|^2] + \mathbb{E}[\|\eta_t \mathbf{g}_{n,t} - \eta_t \bar{\mathbf{g}}_{n,t}\|^2] \\ &\quad - 2\mathbb{E}\langle \mathbf{w}_{n,t} - \mathbf{w}^* - \eta_t \bar{\mathbf{g}}_{n,t}, \eta_t \mathbf{g}_{n,t} - \eta_t \bar{\mathbf{g}}_{n,t} \rangle \\ &\leq (1 - \frac{\mu H \eta_t}{2}) \mathbb{E}\|\mathbf{w}_{n,t} - \mathbf{w}^*\|^2 + 4H\Lambda\eta_t \\ &\quad + \frac{H(H-1)(2H-1)}{3} \ell \eta_t^3 G^2 + \eta_t^2 H^2 (3\delta^2 + 2G^2), \end{aligned} \quad (3)$$

where  $\mathbb{E}\langle \mathbf{w}_{n,t} - \mathbf{w}^* - \eta_t \bar{\mathbf{g}}_{n,t}, \eta_t \mathbf{g}_{n,t} - \eta_t \bar{\mathbf{g}}_{n,t} \rangle = 0$ .

With diminishing learning rate  $\eta_t = \frac{\varrho}{t+\iota}$  with  $\varrho > 0$ , and  $\iota > 0$ , we claim that

$$\mathbb{E}[\|\mathbf{w}_{n,t} - \mathbf{w}^*\|^2] \leq \frac{\zeta}{t+\iota} + \frac{8\Lambda}{\mu}, \quad (4)$$

where  $\zeta = \max\{\ell \mathbb{E}\|\mathbf{w}_0 - \mathbf{w}^*\|^2, \frac{2(H(H-1)(2H-1)\ell G^2 \varrho^3 + 3\varrho^2 H^2 (3\delta^2 + 2G^2))}{3(\mu H \varrho - 2)}\}$ . This can be proved through induction method.

For  $t = 0$ ,  $\mathbb{E}[\|\mathbf{w}_{n,0} - \mathbf{w}^*\|^2] = \mathbb{E}\|\mathbf{w}_0 - \mathbf{w}^*\|^2 = \frac{\ell \mathbb{E}\|\mathbf{w}_0 - \mathbf{w}^*\|^2}{\iota} \leq \frac{\zeta}{\iota} + \frac{8\Lambda}{\mu}$ , where  $\mathbf{w}_{n,0} = \mathbf{w}_0$  denotes the initial global model. For the case of  $t + 1$ ,

$$\begin{aligned} &\mathbb{E}[\|\mathbf{w}_{n,t+1} - \mathbf{w}^*\|^2] \\ &\leq (1 - \frac{\mu H}{2} \frac{\varrho}{t+\iota}) (\frac{\zeta}{t+\iota} + \frac{8\Lambda}{\mu}) + (\frac{\varrho}{t+\iota})^2 H^2 (3\delta^2 + 2G^2) \\ &\quad + \frac{4H\Lambda\varrho}{t+\iota} + \frac{H(H-1)(2H-1)}{3} \ell G^2 (\frac{\varrho}{t+\iota})^3 \end{aligned}$$

$$\begin{aligned}
&= \frac{\zeta}{t+\iota} - \frac{\mu H \varrho}{2(t+\iota)^2} \zeta + \frac{8\Lambda}{\mu} + \frac{\varrho^2 H^2 (3\delta^2 + 2G^2)}{(t+\iota)^2} - \frac{4H\Lambda\varrho}{t+\iota} \\
&+ \frac{4H\Lambda\varrho}{t+\iota} + \frac{H(H-1)(2H-1)\ell G^2 \varrho^3}{3(t+\iota)^3} \\
&= \frac{t+\iota-1}{(t+\iota)^2} \zeta + \frac{8\Lambda}{\mu} + \frac{(t+\iota)(2-\mu H\varrho)}{2(t+\iota)^3} \zeta \\
&+ \frac{H(H-1)(2H-1)\ell G^2 \varrho^3 + 3(t+\iota)\varrho^2 H^2 (3\delta^2 + 2G^2)}{3(t+\iota)^3} \\
&\leq \frac{t+\iota-1}{(t+\iota)^2-1} \zeta + \frac{8\Lambda}{\mu} \\
&\leq \frac{\zeta}{t+\iota+1} + \frac{8\Lambda}{\mu}, \tag{5}
\end{aligned}$$

where  $\zeta$  satisfies  $\zeta \geq \frac{2(H(H-1)(2H-1)\ell G^2 \varrho^3 + 3\varrho^2 H^2 (3\delta^2 + 2G^2))}{3(\mu H \varrho - 2)}$ . By setting  $\varrho > \frac{2}{\mu} > 0$ , so we have

$$\begin{aligned}
\zeta &= \max\{\iota \mathbb{E}\|\mathbf{w}_0 - \mathbf{w}^*\|^2, \frac{J_1(H)}{3(\mu H \varrho - 2)}\} \\
&\leq \iota \mathbb{E}\|\mathbf{w}_0 - \mathbf{w}^*\|^2 + \frac{J_1(H)}{3(\mu H \varrho - 2)}, \tag{6}
\end{aligned}$$

where  $J_1(H) = 2(H(H-1)(2H-1)\ell G^2 \varrho^3 + 3\varrho^2 H^2 (3\delta^2 + 2G^2))$ .

Then applying (5) and (6) into (1), we obtain

$$\begin{aligned}
\mathbb{E}[F(\mathbf{w}_{t+1})] - F^* &\leq \frac{\ell N}{2} \sum_{n=1}^N m_{n,t}^2 \rho_{n,t}^2 \mathbb{E}[\|\mathbf{w}_{n,t+1} - \mathbf{w}^*\|^2] \\
&\leq \frac{\ell N}{2} \sum_{n=1}^N m_{n,t} \rho_{n,t}^2 \left( \frac{\zeta}{t+\iota+1} + \frac{8\Lambda}{\mu} \right) \\
&\leq \frac{\ell N \sum_{n=1}^N m_{n,t} \rho_{n,t}^2}{2(t+\iota+1)} \left( \iota \mathbb{E}\|\mathbf{w}_0 - \mathbf{w}^*\|^2 + \frac{J_1(H)}{3(\mu H \varrho - 2)} \right) \\
&+ \frac{4\ell \Lambda N \sum_{n=1}^N m_{n,t} \rho_{n,t}^2}{\mu}. \tag{7}
\end{aligned}$$

Now, we complete the proof of Theorem 1.

## APPENDIX B PROOF OF THEOREM 2

For any  $t_0 \leq t$ , we have

$$\begin{aligned}
\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_{t_0}\|^2] &\leq N \sum_{n=1}^N \mathbb{E}[\|m_{n,t} \rho_{n,t} (\mathbf{w}_{n,t+1} - \mathbf{w}_{t_0})\|^2] \\
&= N \sum_{n=1}^N m_{n,t}^2 \rho_{n,t}^2 \mathbb{E}[\|\mathbf{w}_{n,t+1} - \mathbf{w}_{t_0}\|^2]. \tag{8}
\end{aligned}$$

Similar to the process of Theorem 1, we have

$$\begin{aligned}
&\mathbb{E}[\|\mathbf{w}_{n,t+1} - \mathbf{w}_{t_0}\|^2] \\
&= \mathbb{E}[\|\mathbf{w}_{n,t} - \eta_t \mathbf{g}_{n,t} - \mathbf{w}_{t_0}\|^2] \\
&= \mathbb{E}[\|\mathbf{w}_{n,t} - \mathbf{w}_{t_0} - \eta_t \mathbf{g}_{n,t} - \eta_t \bar{\mathbf{g}}_{n,t} + \eta_t \bar{\mathbf{g}}_{n,t}\|^2] \\
&= \mathbb{E}[\|\mathbf{w}_{n,t} - \mathbf{w}_{t_0} - \eta_t \bar{\mathbf{g}}_{n,t}\|^2] + \mathbb{E}[\|\eta_t \mathbf{g}_{n,t} - \eta_t \bar{\mathbf{g}}_{n,t}\|^2] \\
&- 2\mathbb{E}\langle \mathbf{w}_{n,t} - \mathbf{w}_{t_0} - \eta_t \bar{\mathbf{g}}_{n,t}, \eta_t \mathbf{g}_{n,t} - \eta_t \bar{\mathbf{g}}_{n,t} \rangle \\
&\leq \mathbb{E}[\|\mathbf{w}_{n,t} - \mathbf{w}_{t_0} - \eta_t \bar{\mathbf{g}}_{n,t}\|^2] + \eta_t H^2 \delta^2. \tag{9}
\end{aligned}$$

Next, we can bound the first term in (9) as follows:

$$\begin{aligned}
&\mathbb{E}\|\mathbf{w}_{n,t} - \mathbf{w}_{t_0} - \eta_t \bar{\mathbf{g}}_{n,t}\|^2 \\
&= \mathbb{E}\|\mathbf{w}_{n,t} - \mathbf{w}_{t_0}\|^2 + \mathbb{E}\|\eta_t \bar{\mathbf{g}}_{n,t}\|^2 - 2\mathbb{E}\langle \mathbf{w}_{n,t} - \mathbf{w}_{t_0}, \eta_t \bar{\mathbf{g}}_{n,t} \rangle \\
&\leq \mathbb{E}\|\mathbf{w}_{n,t} - \mathbf{w}_{t_0}\|^2 + 2\eta_t^2 H^2 (\delta^2 + G^2) \\
&+ 2\eta_t \mathbb{E}\langle \mathbf{w}_{t_0} - \mathbf{w}_{n,t}, \bar{\mathbf{g}}_{n,t} \rangle \\
&\stackrel{(a)}{\leq} \mathbb{E}\|\mathbf{w}_{n,t} - \mathbf{w}_{t_0}\|^2 + 2\eta_t^2 H^2 (\delta^2 + G^2) \\
&+ \frac{\eta_t}{H} \mathbb{E}\|\mathbf{w}_{n,t} - \mathbf{w}_{t_0}\|^2 + \eta_t H \mathbb{E}\|\bar{\mathbf{g}}_{n,t}\|^2 \\
&\leq \mathbb{E}\|\mathbf{w}_{n,t} - \mathbf{w}_{t_0}\|^2 + 2\eta_t^2 H^2 (\delta^2 + G^2) \\
&+ \frac{\eta_t}{H} \mathbb{E}\|\mathbf{w}_{n,t} - \mathbf{w}_{t_0}\|^2 + 2\eta_t H^3 (\delta^2 + G^2) \\
&= (1 + \frac{\eta_t}{H}) \mathbb{E}\|\mathbf{w}_{n,t} - \mathbf{w}_{t_0}\|^2 + 2\eta_t H^2 (\delta^2 + G^2) (\eta_t + H) \tag{10}
\end{aligned}$$

where (a) is by Cauchy-Schwarz inequality and AM-GM inequality.

Plugging (10) back to (9), we obtain

$$\begin{aligned}
\mathbb{E}\|\mathbf{w}_{n,t+1} - \mathbf{w}_{t_0}\|^2 &\leq (1 + \frac{\eta_t}{H}) \mathbb{E}\|\mathbf{w}_{n,t} - \mathbf{w}_{t_0}\|^2 \\
&+ 2\eta_t H^2 (\delta^2 + G^2) (\eta_t + H) + \eta_t H^2 \delta^2. \tag{11}
\end{aligned}$$

After the recursion with  $\eta_t = \eta$ , we further have

$$\begin{aligned}
\mathbb{E}\|\mathbf{w}_{n,t} - \mathbf{w}_{t_0}\|^2 &\leq (1 + \frac{\eta}{H}) \mathbb{E}\|\mathbf{w}_{n,t-1} - \mathbf{w}_{t_0}\|^2 \\
&+ 2\eta H^2 (\delta^2 + G^2) (\eta + H) + \eta H^2 \delta^2 \\
&\leq (1 + \frac{\eta}{H})^2 \mathbb{E}\|\mathbf{w}_{n,t-2} - \mathbf{w}_{t_0}\|^2 + (1 + \frac{\eta}{H}) (\eta H^2 \delta^2 + \\
&2\eta H^2 (\delta^2 + G^2) (\eta + H)) + 2\eta H^2 (\delta^2 + G^2) (\eta + H) + \eta H^2 \delta^2 \\
&\leq \dots \leq (2\eta H^2 (\delta^2 + G^2) (\eta + H) + \eta H^2 \delta^2) \sum_{i=0}^{t-1} (1 + \frac{\eta}{H})^i \\
&\stackrel{(a)}{\leq} (2\eta H^2 (\delta^2 + G^2) (\eta + H) + \eta H^2 \delta^2) (1 + \frac{\eta}{H})^{t-1}, \tag{12}
\end{aligned}$$

where (a) follows the summation of geometric progression, we have  $\sum_{i=0}^{t-1} (1 + \frac{\eta}{H})^i = \frac{(1 + \frac{\eta}{H})^t - 1}{\frac{\eta}{H}} \leq (1 + \frac{\eta}{H})^{t-1}$ .

Hence, (11) can be rewritten as

$$\begin{aligned}
\mathbb{E}\|\mathbf{w}_{n,t+1} - \mathbf{w}_{t_0}\|^2 &\leq 2\eta H^2 (\delta^2 + G^2) (\eta + H) + \eta H^2 \delta^2 \\
&+ (2\eta H^2 (\delta^2 + G^2) (\eta + H) + \eta H^2 \delta^2) (1 + \frac{\eta}{H})^t. \tag{13}
\end{aligned}$$

For  $t_0 = t$ , we achieve

$$\begin{aligned}
\mathbb{E}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 &\leq N M_t (2\eta H^2 (\delta^2 + G^2) (\eta + H) + \eta H^2 \delta^2 \\
&+ (2\eta H^2 (\delta^2 + G^2) (\eta + H) + \eta H^2 \delta^2) (1 + \frac{\eta}{H})^t), \tag{14}
\end{aligned}$$

where  $M_t = \sum_{n=1}^N m_{n,t} \rho_{n,t}^2$ .

Given Assumption 1, the global loss function  $F(\mathbf{w})$  is  $\ell$ -smooth, so we have

$$\begin{aligned}
&\mathbb{E}[F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t)] \\
&\leq \mathbb{E}\langle \nabla F(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t \rangle + \frac{\ell}{2} \mathbb{E}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 \\
&= \mathbb{E}\langle \nabla F(\mathbf{w}_t), \sum_{n=1}^N m_{n,t} \rho_{n,t} (\mathbf{w}_{n,t} - \eta \mathbf{g}_{n,t} - \mathbf{w}_t) \rangle
\end{aligned}$$

$$\begin{aligned}
& + \frac{\ell}{2} \mathbb{E} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 \\
& \leq -\eta \mathbb{E} \langle \nabla F(\mathbf{w}_t), \mathbf{g}_{n,t} \rangle + \mathbb{E} \langle \nabla F(\mathbf{w}_t), \mathbf{w}_{n,t} - \mathbf{w}_t \rangle \\
& + \frac{\ell}{2} \mathbb{E} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 \\
& \stackrel{(a)}{=} -\frac{\eta}{2} \mathbb{E} [\|\nabla F(\mathbf{w}_t)\|^2 + \|\mathbf{g}_{n,t}\|^2 - \|\nabla F(\mathbf{w}_t) - \mathbf{g}_{n,t}\|^2] \\
& + \mathbb{E} \langle \nabla F(\mathbf{w}_t), \mathbf{w}_{n,t} - \mathbf{w}_t \rangle + \frac{\ell}{2} \mathbb{E} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 \\
& \leq -\frac{\eta}{2} \mathbb{E} [\|\nabla F(\mathbf{w}_t)\|^2] + \frac{\eta}{2} \mathbb{E} [\|\nabla F(\mathbf{w}_t) - \mathbf{g}_{n,t}\|^2] \\
& + \mathbb{E} \langle \nabla F(\mathbf{w}_t), \mathbf{w}_{n,t} - \mathbf{w}_t \rangle + \frac{\ell}{2} \mathbb{E} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2 \\
& \stackrel{(b)}{\leq} -\frac{\eta}{2} \mathbb{E} [\|\nabla F(\mathbf{w}_t)\|^2] + \frac{\eta \sigma^2}{2} + \frac{1}{2D} \mathbb{E} [\|\nabla F(\mathbf{w}_t)\|^2] \\
& + \frac{D}{2} \mathbb{E} [\|\mathbf{w}_{n,t} - \mathbf{w}_t\|^2] + \frac{\ell}{2} \mathbb{E} [\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2] \\
& = (\frac{1}{2D} - \frac{\eta}{2}) \mathbb{E} [\|\nabla F(\mathbf{w}_t)\|^2] + \frac{\eta \sigma^2}{2} + \frac{D}{2} \mathbb{E} [\|\mathbf{w}_{n,t} - \mathbf{w}_t\|^2] \\
& + \frac{\ell}{2} \mathbb{E} [\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2], \tag{15}
\end{aligned}$$

where (a) follows  $\langle \mathbf{a}, \mathbf{b} \rangle = \frac{1}{2} \|\mathbf{a}\|^2 + \frac{1}{2} \|\mathbf{b}\|^2 - \frac{1}{2} \|\mathbf{a} - \mathbf{b}\|^2$ , (b) follows Cauchy-Schwarz inequality and AM-GM inequality, and  $D > 0$  is a constant.

Applying (14) and (12) to (15), we obtain

$$\begin{aligned}
& \mathbb{E}[F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t)] \\
& \leq (\frac{1}{2D} - \frac{\eta}{2}) \mathbb{E} [\|\nabla F(\mathbf{w}_t)\|^2] + \frac{\eta \sigma^2}{2} + \frac{J_2(H)D}{2} (1 + \frac{\eta}{H})^{t-1} \\
& + \frac{\ell J_2(H)N}{2} M_t (1 + \frac{\eta}{H})^t + \frac{\ell J_2(H)NM_t}{2}, \tag{16}
\end{aligned}$$

where  $J_2(H) = 2\eta H^2(\delta^2 + G^2)(\eta + H) + \eta H^2 \delta^2$ .

Suppose  $\eta > \frac{1}{D}$  and recursively utilizing the above inequality from  $t = 0$  to  $t = T - 1$ , we obtain

$$\begin{aligned}
& \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\nabla F(\mathbf{w}_t)\|^2 \\
& \leq \frac{\mathbb{E}[F(\mathbf{w}_0)] - \mathbb{E}[F(\mathbf{w}_T)]}{T(\frac{\eta}{2} - \frac{1}{2D})} + \frac{J_2(H)((1 + \frac{\eta}{H})^{T-1} - \frac{1}{1 + \frac{\eta}{H}})}{T \frac{\eta}{H} (\frac{\eta}{D} - \frac{1}{D^2})} \\
& + \frac{\ell J_2(H)N \sum_{t=0}^{T-1} M_t (1 + \frac{\eta}{H})^t}{T(\eta - \frac{1}{D})} + \frac{\ell J_2(H)N \sum_{t=0}^{T-1} M_t}{T(\eta - \frac{1}{D})} \\
& + \frac{\eta \sigma^2}{\eta - \frac{1}{D}}. \tag{17}
\end{aligned}$$

By setting  $\eta < H(e^{\frac{1}{T}} - 1)$ , we can conclude that  $\frac{(1 + \frac{\eta}{H})^T}{T}$  is a monotonic decreasing function. Therefore, when  $T \rightarrow \infty$ , the first, second, third and fourth terms on the right-hand side of (17) go to zero. To this end, Theorem 2 is proved.

#### APPENDIX C PROOF OF LEMMA 3

We first split the following term into three terms as follows:

$$\begin{aligned}
\mathbb{E} \|\mathbf{w}_{n,t} - \mathbf{w}^* - \eta_t \bar{\mathbf{g}}_{n,t}\|^2 & = \mathbb{E} \|\mathbf{w}_{n,t} - \mathbf{w}^*\|^2 + \mathbb{E} [\|\eta_t \bar{\mathbf{g}}_{n,t}\|^2] \\
& - 2\mathbb{E} \langle \mathbf{w}_{n,t} - \mathbf{w}^*, \eta_t \bar{\mathbf{g}}_{n,t} \rangle. \tag{18}
\end{aligned}$$

The second term in (18) can be bounded as follows:

$$\begin{aligned}
& \mathbb{E} [\|\eta_t \bar{\mathbf{g}}_{n,t}\|^2] \\
& \stackrel{(a)}{\leq} 2\eta_t^2 \mathbb{E} [\|\bar{\mathbf{g}}_{n,t} - \mathbf{g}_{n,t}\|^2] + 2\eta_t^2 \mathbb{E} [\|\mathbf{g}_{n,t}\|^2] \\
& = 2\eta_t^2 \mathbb{E} [\|\sum_{\tau=0}^{H-1} (\nabla F_n(\mathbf{w}_{n,t}^\tau) - \nabla F_n(\mathbf{w}_{n,t}^\tau; \mathcal{B}_{n,t}^\tau))\|^2] \\
& + 2\eta_t^2 \mathbb{E} [\|\sum_{\tau=0}^{H-1} \nabla F_n(\mathbf{w}_{n,t}^\tau; \mathcal{B}_{n,t}^\tau)\|^2] \\
& \stackrel{(b)}{\leq} 2\eta_t^2 H^2 \delta^2 + 2\eta_t^2 H^2 G^2 \\
& = 2\eta_t^2 H^2 (\delta^2 + G^2), \tag{19}
\end{aligned}$$

where (a) follows by applying  $\|\mathbf{a}\|^2 \leq 2\|\mathbf{a} - \mathbf{b}\|^2 + 2\|\mathbf{b}\|^2$ , and (b) is by Jensen's inequality, Assumptions 3 and 4.

Next, we can bound the third term in (18) as follows:

$$\begin{aligned}
& -2\mathbb{E} \langle \mathbf{w}_{n,t} - \mathbf{w}^*, \eta_t \bar{\mathbf{g}}_{n,t} \rangle \\
& = -2\eta_t \sum_{\tau=0}^{H-1} \mathbb{E} \langle \mathbf{w}_{n,t} - \mathbf{w}^*, \nabla F_n(\mathbf{w}_{n,t}^\tau) \rangle \\
& \stackrel{(a)}{\leq} -2\eta_t \sum_{\tau=0}^{H-1} \mathbb{E} [(F_n(\mathbf{w}_{n,t}) - F_n(\mathbf{w}^*) + \frac{\mu}{4} \|\mathbf{w}_{n,t} - \mathbf{w}^*\|^2 \\
& - \ell \|\mathbf{w}_{n,t}^\tau - \mathbf{w}_{n,t}\|^2)] \\
& = -2\eta_t \sum_{\tau=0}^{H-1} \mathbb{E} [(F_n(\mathbf{w}_{n,t}) - F^*) - (F_n(\mathbf{w}^*) - F^*) \\
& + \frac{\mu}{4} \|\mathbf{w}_{n,t} - \mathbf{w}^*\|^2 - \ell \|\mathbf{w}_{n,t}^\tau - \mathbf{w}_{n,t}\|^2] \\
& \leq -2\eta_t \sum_{\tau=0}^{H-1} \mathbb{E} [(F_n(\mathbf{w}_{n,t}) - F^*) - |F_n(\mathbf{w}^*) - F^*| \\
& + \frac{\mu}{4} \|\mathbf{w}_{n,t} - \mathbf{w}^*\|^2 - \ell \|\mathbf{w}_{n,t}^\tau - \mathbf{w}_{n,t}\|^2] \\
& \leq 2\eta_t \sum_{\tau=0}^{H-1} \mathbb{E} [(F^* - F_n^*) + \Lambda - \frac{\mu}{4} \|\mathbf{w}_{n,t} - \mathbf{w}^*\|^2 \\
& + \ell \|\mathbf{w}_{n,t}^\tau - \mathbf{w}_{n,t}\|^2] \\
& \leq 2\eta_t \sum_{\tau=0}^{H-1} \mathbb{E} [|F^* - F_n^*| + \Lambda - \frac{\mu}{4} \|\mathbf{w}_{n,t} - \mathbf{w}^*\|^2 \\
& + \ell \|\mathbf{w}_{n,t}^\tau - \mathbf{w}_{n,t}\|^2] \\
& \leq 4\eta_t H \Lambda - \frac{\mu \eta_t}{2} \sum_{\tau=0}^{H-1} \mathbb{E} [\|\mathbf{w}_{n,t} - \mathbf{w}^*\|^2] \\
& + 2\eta_t \ell \sum_{\tau=0}^{H-1} \mathbb{E} [\|\mathbf{w}_{n,t}^\tau - \mathbf{w}_{n,t}\|^2] \\
& = 4\eta_t H \Lambda - \frac{\mu \eta_t H}{2} \mathbb{E} [\|\mathbf{w}_{n,t} - \mathbf{w}^*\|^2] \\
& + 2\eta_t \ell \sum_{\tau=0}^{H-1} \mathbb{E} [\|\sum_{j=0}^{\tau-1} \mathbf{g}_{n,t}^j\|^2] \\
& \stackrel{(b)}{\leq} 4\eta_t H \Lambda - \frac{\mu \eta_t H}{2} \mathbb{E} [\|\mathbf{w}_{n,t} - \mathbf{w}^*\|^2] \\
& + \frac{H(H-1)(2H-1)}{3} \ell \eta_t^3 G^2. \tag{20}
\end{aligned}$$

where (a) is due to the Lemma 1, and (b) follows the Jensen's inequality and  $\sum_{\tau=0}^{H-1} \tau^2 = \frac{H(H-1)(2H-1)}{6}$ .

Combining (19) and (20), we can conclude that

$$\begin{aligned} & \mathbb{E} \|\mathbf{w}_{n,t} - \mathbf{w}^* - \eta_t \bar{\mathbf{g}}_{n,t}\|^2 \\ & \leq (1 - \frac{\mu H \eta_t}{2}) \mathbb{E} \|\mathbf{w}_{n,t} - \mathbf{w}^*\|^2 + 4H\Lambda\eta_t \\ & \quad + \frac{H(H-1)(2H-1)}{3} \ell \eta_t^3 G^2 + 2\eta_t^2 H^2 (\delta^2 + G^2). \end{aligned} \quad (21)$$

Now, we complete the proof of Lemma 3.