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 $\boldsymbol{g}_{n,t} = \sum_{\tau=0}^{H-1} \boldsymbol{g}_{n,t}^{\tau}$ and $\bar{\boldsymbol{g}}_{n,t} = \sum_{\tau=0}^{H-1} \bar{\boldsymbol{g}}_{n,t}^{\tau}$, thus $\mathbb{E}[\boldsymbol{g}_{n,t}] = \bar{\boldsymbol{g}}_{n,t}$. 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 ℓ-smooth, we have

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

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 $\|\sum_{n=1}^{N}\boldsymbol{a}_n\|^2 \le N\sum_{n=1}^{N}\|\boldsymbol{a}_n\|^2$. Now, we can bound $\mathbb{E}[\|\mathbf{w}_{n,t+1} - \mathbf{w}^*\|^2]$ by applying the following lemmas.

Lemma 2. *Under Assumptions 1 and 2, for any $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{z} \in \mathbb{R}^d$, we have $<\nabla F_n(\boldsymbol{x}), \boldsymbol{z} - \boldsymbol{y}> \ge F_n(\boldsymbol{z}) - F_n(\boldsymbol{y}) + \frac{\mu}{4}\|\boldsymbol{y} - \boldsymbol{z}\|^2 - \ell\|\boldsymbol{z} - \boldsymbol{x}\|^2$, $\forall n$ [1].*

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

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

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

where (a) follows by the $\|\sum_{\tau=0}^{H-1}\boldsymbol{a}_{\tau}\|^2 \le H\sum_{\tau=0}^{H-1}\|\boldsymbol{a}_{\tau}\|^2$, and (b) follows Assumption 3.

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

Proof. The proof is presented in Appendix C. □

Hence, $\mathbb{E}[\|\mathbf{w}_{n,t+1} - \mathbf{w}^*\|^2]$ with Lemmas 3 and 4 can be rewritten as

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

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

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

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

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

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

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

H. Ao, H. Tian, and G. Nie are with the 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 the 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).

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

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

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

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

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

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

Now, we complete the proof of Theorem 1.

Appendix B
Proof of Theorem 2

For any $t_0 \le t$, we have

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

Similar to the process of Theorem 1, we have

$$\mathbb{E}[\|\mathbf{w}_{n,t+1}-\mathbf{w}_{t_0}\|^2]$$
$$= \mathbb{E}[\|\mathbf{w}_{n,t}-\eta_t\boldsymbol{g}_{n,t}-\mathbf{w}_{t_0}\|^2]$$
$$= \mathbb{E}[\|\mathbf{w}_{n,t}-\mathbf{w}_{t_0}-\eta_t\boldsymbol{g}_{n,t}-\eta_t\bar{\boldsymbol{g}}_{n,t}+\eta_t\bar{\boldsymbol{g}}_{n,t}\|^2]$$
$$= \mathbb{E}[\|\mathbf{w}_{n,t}-\mathbf{w}_{t_0}-\eta_t\bar{\boldsymbol{g}}_{n,t}\|^2] + \mathbb{E}[\|\eta_t\boldsymbol{g}_{n,t}-\eta_t\bar{\boldsymbol{g}}_{n,t}\|^2]$$
$$- 2\mathbb{E}<\mathbf{w}_{n,t}-\mathbf{w}_{t_0}-\eta_t\bar{\boldsymbol{g}}_{n,t}, \eta_t\boldsymbol{g}_{n,t}-\eta_t\bar{\boldsymbol{g}}_{n,t}>$$
$$\le \mathbb{E}[\|\mathbf{w}_{n,t}-\mathbf{w}_{t_0}-\eta_t\bar{\boldsymbol{g}}_{n,t}\|^2] + \eta_t H^2\delta^2. \tag{9}$$

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

$$\mathbb{E}\|\mathbf{w}_{n,t}-\mathbf{w}_{t_0}-\eta_t\bar{\boldsymbol{g}}_{n,t}\|^2$$
$$= \mathbb{E}\|\mathbf{w}_{n,t}-\mathbf{w}_{t_0}\|^2 + \mathbb{E}\|\eta_t\bar{\boldsymbol{g}}_{n,t}\|^2 - 2\mathbb{E}<\mathbf{w}_{n,t}-\mathbf{w}_{t_0}, \eta_t\bar{\boldsymbol{g}}_{n,t}>$$
$$\le \mathbb{E}\|\mathbf{w}_{n,t}-\mathbf{w}_{t_0}\|^2 + 2\eta_t^2H^2(\delta^2+G^2)$$
$$+ 2\eta_t\mathbb{E}<\mathbf{w}_{t_0}-\mathbf{w}_{n,t}, \bar{\boldsymbol{g}}_{n,t}>$$
$$\overset{(a)}{\le} \mathbb{E}\|\mathbf{w}_{n,t}-\mathbf{w}_{t_0}\|^2 + 2\eta_t^2H^2(\delta^2+G^2)$$
$$+ \frac{\eta_t}{H}\mathbb{E}\|\mathbf{w}_{n,t}-\mathbf{w}_{t_0}\|^2 + \eta_t H\mathbb{E}\|\bar{\boldsymbol{g}}_{n,t}\|^2$$
$$\le \mathbb{E}\|\mathbf{w}_{n,t}-\mathbf{w}_{t_0}\|^2 + 2\eta_t^2H^2(\delta^2+G^2)$$
$$+ \frac{\eta_t}{H}\mathbb{E}\|\mathbf{w}_{n,t}-\mathbf{w}_{t_0}\|^2 + 2\eta_t H^3(\delta^2+G^2)$$
$$= (1+\frac{\eta_t}{H})\mathbb{E}\|\mathbf{w}_{n,t}-\mathbf{w}_{t_0}\|^2 + 2\eta_t H^2(\delta^2+G^2)(\eta_t+H) \tag{10}$$

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

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

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

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

$$\mathbb{E}\|\mathbf{w}_{n,t}-\mathbf{w}_{t_0}\|^2 \le (1+\frac{\eta}{H})\mathbb{E}\|\mathbf{w}_{n,t-1}-\mathbf{w}_{t_0}\|^2$$
$$+ 2\eta H^2(\delta^2+G^2)(\eta+H) + \eta H^2\delta^2$$
$$\le (1+\frac{\eta}{H})^2\mathbb{E}\|\mathbf{w}_{n,t-2}-\mathbf{w}_{t_0}\|^2 + (1+\frac{\eta}{H})(\eta H^2\delta^2 +$$
$$2\eta_t H^2(\delta^2+G^2)(\eta+H)) + 2\eta H^2(\delta^2+G^2)(\eta+H) + \eta H^2\delta^2$$
$$\le \dots \le (2\eta H^2(\delta^2+G^2)(\eta+H) + \eta H^2\delta^2)\sum_{i=0}^{t-1}(1+\frac{\eta}{H})^i$$
$$\overset{(a)}{\le} (2\eta H^2(\delta^2+G^2)(\eta+H) + \eta H^2\delta^2)(1+\frac{\eta}{H})^{t-1}, \tag{12}$$

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

Hence, (11) can be rewritten as

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

For $t_0 = t$, we achieve

$$\mathbb{E}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|^2 \le NM_t(2\eta H^2(\delta^2+G^2)(\eta+H) + \eta H^2\delta^2$$
$$+ (2\eta H^2(\delta^2+G^2)(\eta+H) + \eta H^2\delta^2)(1+\frac{\eta}{H})^t), \tag{14}$$

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 ℓ-smooth, so we have

$$\mathbb{E}[F(\mathbf{w}_{t+1}) - F(\mathbf{w}_t)]$$
$$\le \mathbb{E}<\nabla F(\mathbf{w}_t), \mathbf{w}_{t+1}-\mathbf{w}_t> + \frac{\ell}{2}\mathbb{E}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|^2$$
$$= \mathbb{E}<\nabla F(\mathbf{w}_t), \sum_{n=1}^{N} m_{n,t}\rho_{n,t}(\mathbf{w}_{n,t}-\eta\boldsymbol{g}_{n,t}-\mathbf{w}_t)>$$

$$+\frac{\ell}{2}\mathbb{E}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|^2$$
$$\leq -\eta\mathbb{E}<\nabla F(\mathbf{w}_t), \boldsymbol{g}_{n,t}> + \mathbb{E}<\nabla F(\mathbf{w}_t), \mathbf{w}_{n,t}-\mathbf{w}_t>$$
$$+\frac{\ell}{2}\mathbb{E}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|^2$$
$$\overset{(a)}{=} -\frac{\eta}{2}\mathbb{E}[\|\nabla F(\mathbf{w}_t)\|^2 + \|\boldsymbol{g}_{n,t}\|^2 - \|\nabla F(\mathbf{w}_t)-\boldsymbol{g}_{n,t}\|^2]$$
$$+\mathbb{E}<\nabla F(\mathbf{w}_t), \mathbf{w}_{n,t}-\mathbf{w}_t> + \frac{\ell}{2}\mathbb{E}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|^2$$
$$\leq -\frac{\eta}{2}\mathbb{E}[\|\nabla F(\mathbf{w}_t)\|^2] + \frac{\eta}{2}\mathbb{E}[\|\nabla F(\mathbf{w}_t)-\boldsymbol{g}_{n,t}\|^2]$$
$$+\mathbb{E}<\nabla F(\mathbf{w}_t), \mathbf{w}_{n,t}-\mathbf{w}_t> + \frac{\ell}{2}\mathbb{E}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|^2$$
$$\overset{(b)}{\leq} -\frac{\eta}{2}\mathbb{E}[\|\nabla F(\mathbf{w}_t)\|^2] + \frac{\eta\sigma^2}{2} + \frac{1}{2D}\mathbb{E}[\|\nabla F(\mathbf{w}_t)\|^2]$$
$$+\frac{D}{2}\mathbb{E}[\|\mathbf{w}_{n,t}-\mathbf{w}_t\|^2] + \frac{\ell}{2}\mathbb{E}[\|\mathbf{w}_{t+1}-\mathbf{w}_t\|^2]$$
$$=(\frac{1}{2D}-\frac{\eta}{2})\mathbb{E}[\|\nabla F(\mathbf{w}_t)\|^2] + \frac{\eta\sigma^2}{2} + \frac{D}{2}\mathbb{E}[\|\mathbf{w}_{n,t}-\mathbf{w}_t\|^2]$$
$$+\frac{\ell}{2}\mathbb{E}[\|\mathbf{w}_{t+1}-\mathbf{w}_t\|^2], \tag{15}$$

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

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

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

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

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

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 $T\to\infty$, 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 4

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

$$\mathbb{E}\|\mathbf{w}_{n,t}-\mathbf{w}^*-\eta_t\bar{\boldsymbol{g}}_{n,t}\|^2 = \mathbb{E}\|\mathbf{w}_{n,t}-\mathbf{w}^*\|^2 + \mathbb{E}[\|\eta_t\bar{\boldsymbol{g}}_{n,t}\|^2]$$
$$-2\mathbb{E}<\mathbf{w}_{n,t}-\mathbf{w}^*, \eta_t\bar{\boldsymbol{g}}_{n,t}>. \tag{18}$$

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

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

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

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

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

where (a) is due to the Lemma 2, 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

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

Now, we complete the proof of Lemma 4.

References

[1] S. P. Karimireddy, S. Kale, M. Mohri, S. J. Reddi, S. U. Stich, and A. T. Suresh, "SCAFFOLD: Stochastic controlled averaging for federated learning," in *Proc. ICML*, vol. 119, Jul. 2020, pp. 5132–5143.