Supplementary Material

Junhe Zhang, Wanli Ni, Dongyu Wang

Appendix

A. Proof of Lemma 2

Fix training round $t \geq 1$. Considering the largest $t_0 \leq t$ that satisfies $t_0 \bmod I = 0$ (Note that such t_0 must exist and $t - t_0 \leq I$.) Recalling $\mathbf{w}_{\mathbf{c},k,t+1} = \tilde{\mathbf{w}}_{\mathbf{c},k,t} - \eta\tilde{\mathbf{g}}'_{\mathbf{c},k,t}$ and $\mathbf{w}_{\mathbf{c},t+1} = \frac{1}{K}\sum_{k=1}^{K}\mathbf{w}_{\mathbf{c},k,t+1}$ for client-side model updating and aggregation, using $\mathbf{m}_t$ to represent the binary matrix obtained by aggregating $\mathbf{m}_{k,t}$ and performing element-wise normalization, we have

$$\mathbf{w}_{\mathbf{c},k,t} = \mathbf{m}_{k,t} \odot (\mathbf{w}_{\mathbf{c},t_0} - \eta\sum_{\tau=t_0}^{t}\mathbf{g}'_{\mathbf{c},k,\tau}) \tag{1}$$

and

$$\mathbf{w}_{\mathbf{c},t} = \mathbf{m}_t \odot \mathbf{w}_{\mathbf{c},t_0} - \eta\sum_{\tau=t_0}^{t}\frac{1}{K}\sum_{k=1}^{K}\mathbf{m}_{k,t}\odot\mathbf{g}'_{\mathbf{c},k,\tau}. \tag{2}$$

where Eqn. (1) follows from

$$\begin{aligned}\mathbf{w}_{\mathbf{c},k,t} =&(((\mathbf{w}_{\mathbf{c},k,t_0} - \eta\mathbf{g}'_{\mathbf{c},k,t_0})\odot\mathbf{m}_{k,t_0} - \eta\mathbf{g}'_{\mathbf{c},k,t_1})\\ &\odot\mathbf{m}_{k,t_1} - \eta\mathbf{g}'_{\mathbf{c},k,t_2})\odot ... \odot\mathbf{m}_{k,t} - \eta\mathbf{g}'_{\mathbf{c},k,t}\odot\mathbf{m}_{k,t}\\ =&\mathbf{w}_{\mathbf{c},k,t_0}\odot\mathbf{m}_{k,t_0}\odot\mathbf{m}_{k,t_1}\odot ... \odot\mathbf{m}_{k,t}\\ &-\eta\mathbf{g}'_{\mathbf{c},k,t_0}\odot\mathbf{m}_{k,t_0}\odot\mathbf{m}_{k,t_1}\odot ... \odot\mathbf{m}_{k,t}\\ &-\eta\mathbf{g}'_{\mathbf{c},k,t_1}\odot\mathbf{m}_{k,t_1}\odot\mathbf{m}_{k,t_2}\odot ... \odot\mathbf{m}_{k,t}\\ &...\\ &-\eta\mathbf{g}'_{\mathbf{c},k,t}\odot\mathbf{m}_{k,t}\\ \overset{(a)}{=}&\mathbf{m}_{k,t}\odot(\mathbf{w}_{\mathbf{c},t_0} - \eta\sum_{\tau=t_0}^{t}\mathbf{g}'_{\mathbf{c},k,\tau}),\end{aligned}$$

where (a) follows from $\mathbf{m}_{k,t_0}\odot\mathbf{m}_{k,t_1}\odot ... \odot\mathbf{m}_{k,t} = \mathbf{m}_{k,t_1}\odot\mathbf{m}_{k,t_2}\odot ... \odot\mathbf{m}_{k,t} = ... = \mathbf{m}_{k,t}$.

Thus, we have

$$\begin{aligned}&\mathbb{E}\|\mathbf{w}_{\mathbf{c},t} - \tilde{\mathbf{w}}_{\mathbf{c},k,t}\|^2\\ =&\mathbb{E}\|\mathbf{w}_{\mathbf{c},t} - \mathbf{w}_{\mathbf{c},k,t} + \mathbf{w}_{\mathbf{c},k,t} - \tilde{\mathbf{w}}_{\mathbf{c},k,t}\|^2\\ \overset{(a)}{\leq}&2\mathbb{E}\|\mathbf{w}_{\mathbf{c},t} - \mathbf{w}_{\mathbf{c},k,t}\|^2 + 2\mathbb{E}\|\mathbf{w}_{\mathbf{c},k,t} - \tilde{\mathbf{w}}_{\mathbf{c},k,t}\|^2\\ \overset{(b)}{\leq}&2\mathbb{E}\|(\mathbf{m}_t - \mathbf{m}_{k,t})\odot\mathbf{w}_{\mathbf{c},t} - \eta(\sum_{\tau=t_0}^{t}\frac{1}{K}\sum_{k=1}^{K}\mathbf{m}_{k,t}\odot\mathbf{g}'_{\mathbf{c},k,\tau}\\ &-\sum_{\tau=t_0}^{t}\mathbf{m}_{k,t}\odot\mathbf{g}'_{\mathbf{c},k,\tau})\|^2 + 2\rho_t\sum_{l=1}^{L_c}W_l^2\\ \overset{(c)}{\leq}&4\eta^2\mathbb{E}\|\sum_{\tau=t_0}^{t}(\frac{1}{K}\sum_{k=1}^{K}\hat{\mathbf{g}}'_{\mathbf{c},k,\tau}) - \sum_{\tau=t_0}^{t}\hat{\mathbf{g}}'_{\mathbf{c},k,\tau}\|^2\\ &+4\mathbb{E}\|(\mathbf{m}_t - \mathbf{m}_{k,t})\odot\mathbf{w}_{\mathbf{c},t}\|^2 + 2\rho_t\sum_{l=1}^{L_c}W_l^2\\ \overset{(d)}{\leq}&4\eta^2\mathbb{E}\|\sum_{\tau=t_0}^{t}(\frac{1}{K}\sum_{k=1}^{K}\hat{\mathbf{g}}'_{\mathbf{c},k,\tau}) - \sum_{\tau=t_0}^{t}\hat{\mathbf{g}}'_{\mathbf{c},k,\tau}\|^2\\ &+4\sum_{l=1}^{L}W_l^2 + 2\rho_t\sum_{l=1}^{L_c}W_l^2,\end{aligned}$$

where $\hat{\mathbf{g}}'_{\mathbf{c},k,\tau} = \mathbf{m}_{k,t}\odot\mathbf{g}'_{\mathbf{c},k,\tau}$ and (a) (c) (d) follows by using the inequality $\|\sum_{i=1}^{n}\mathbf{z}_i\|^2 \leq n\sum_{i=1}^{n}\|\mathbf{z}_i\|^2$ for any vectors $\mathbf{z}_i$ and any positive integer n (using $n = 2$ in (a) and (c), $n = K$ in (d)). (b) follows from Eqn. (1), Eqn. (2) and Assumption 4.

Note that

$$\begin{aligned}&\mathbb{E}\|\sum_{\tau=t_0}^{t}(\frac{1}{K}\sum_{k=1}^{K}\hat{\mathbf{g}}'_{\mathbf{c},k,\tau}) - \sum_{\tau=t_0}^{t}\hat{\mathbf{g}}'_{\mathbf{c},k,\tau}\|^2\\ \overset{(a)}{\leq}&2\mathbb{E}\{\|\sum_{\tau=t_0}^{t}\frac{1}{K}\sum_{k=1}^{K}\hat{\mathbf{g}}'_{\mathbf{c},k,\tau}\|^2 + \|\sum_{\tau=t_0}^{t}\hat{\mathbf{g}}'_{\mathbf{c},k,\tau}\|^2\}\\ \overset{(b)}{\leq}&2(t-t_0+1)\mathbb{E}\{\sum_{\tau=t_0}^{t}\|\frac{1}{K}\sum_{k=1}^{K}\hat{\mathbf{g}}'_{\mathbf{c},k,\tau}\|^2 + \sum_{\tau=t_0}^{t}\|\hat{\mathbf{g}}'_{\mathbf{c},k,\tau}\|^2\}\\ \overset{(c)}{\leq}&2(t-t_0+1)\mathbb{E}\{\sum_{\tau=t_0}^{t}\frac{1}{K}\sum_{k=1}^{K}\|\hat{\mathbf{g}}'_{\mathbf{c},k,\tau}\|^2 + \sum_{\tau=t_0}^{t}\|\hat{\mathbf{g}}'_{\mathbf{c},k,\tau}\|^2\}\\ \overset{(d)}{\leq}&2(I+1)^2\sum_{l=1}^{L_c}G_l^2,\end{aligned}$$

where, (a)-(c) follows by using the inequality $\|\sum_{i=1}^{n}\mathbf{z}_i\|^2 \leq n\sum_{i=1}^{n}\|\mathbf{z}_i\|^2$ and $n = 2$ for (a), $n = (t - t_0 + 1)$ for (b), and $n = K$ for (c); (d) follows from the Assumption 4.

Thus, we have

$$\mathbb{E}\|\mathbf{w}_{\mathbf{c},t} - \tilde{\mathbf{w}}_{\mathbf{c},k,t}\|^2 \leq 8\eta^2(I+1)^2\sum_{l=1}^{L_c}G_l^2 + 4\sum_{l=1}^{L}W_l^2 + 2\rho_t\sum_{l=1}^{L_c}W_l^2.$$

B. Proof of the Theorem 1

For training round $t \leq 1$. By the smoothness of loss function F, we have

$$\begin{aligned}\mathbb{E}[F(\mathbf{w}_{t+1})] \leq& \mathbb{E}[F(\mathbf{w}_t)] + \mathbb{E}[\langle\nabla F(\mathbf{w}_t), \mathbf{w}_{t+1} - \mathbf{w}_t\rangle]\\ &+\frac{\beta}{2}\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2]\end{aligned} \tag{3}$$

Note that

$$\mathbb{E}\|\mathbf{w}_{t+1} - \mathbf{w}_t\|^2$$

Junhe Zhang is with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, China (e-mail: jhzhangbupt@163.com).

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

Dongyu Wang is with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, China (e-mail: dy_wang@bupt.edu.cn).

$$
\begin{aligned}
&=\mathbb{E}\|[\mathbf{w}_{\mathbf{c},t+1};\mathbf{w}_{\mathbf{s},t+1}]-[\mathbf{w}_{\mathbf{c},t};\mathbf{w}_{\mathbf{s},t}]\|^2\\
&=\mathbb{E}\|\mathbf{w}_{\mathbf{c},t+1}-\mathbf{w}_{\mathbf{c},t};\mathbf{w}_{\mathbf{s},t+1}-\mathbf{w}_{\mathbf{s},t}\|^2\\
&=\mathbb{E}\|\mathbf{w}_{\mathbf{c},t+1}-\mathbf{w}_{\mathbf{c},t}\|^2+\mathbb{E}\|\mathbf{w}_{\mathbf{s},t+1}-\mathbf{w}_{\mathbf{s},t}\|^2. \qquad (4)
\end{aligned}
$$

where $\mathbb{E}\|\mathbf{w}_{\mathbf{c},t+1}-\mathbf{w}_{\mathbf{c},t}\|^2$ can be bounded as

$$
\begin{aligned}
&\mathbb{E}\|\mathbf{w}_{\mathbf{c},t+1}-\mathbf{w}_{\mathbf{c},t}\|^2=\mathbb{E}\|\frac{1}{K}\sum_{k=1}^{K}(\mathbf{w}_{\mathbf{c},k,t+1}-\mathbf{w}_{\mathbf{c},k,t})\|^2\\
&=\frac{1}{K^2}\sum_{k=1}^{K}\mathbb{E}\|(\mathbf{w}_{\mathbf{c},k,t+1}-\tilde{\mathbf{w}}_{\mathbf{c},k,t})+(\tilde{\mathbf{w}}_{\mathbf{c},k,t}-\mathbf{w}_{\mathbf{c},k,t})\|^2\\
&\le\frac{2}{K^2}\sum_{k=1}^{K}(\mathbb{E}\|\mathbf{w}_{\mathbf{c},k,t+1}-\tilde{\mathbf{w}}_{\mathbf{c},k,t}\|^2+\mathbb{E}\|\tilde{\mathbf{w}}_{\mathbf{c},k,t}-\mathbf{w}_{\mathbf{c},k,t}\|^2)\\
&\overset{(a)}{\le}\frac{2}{K^2}\sum_{k=1}^{K}(\eta^2\mathbb{E}\|\tilde{\mathbf{g}}'_{\mathbf{c},k,t}\|^2+\rho_t\sum_{l=1}^{L_c}W_l^2)\\
&\overset{(b)}{\le}\frac{2}{K}\sum_{l=1}^{L_c}(\eta^2\sigma_l^2+\rho_tW_l^2)+2\eta^2\mathbb{E}\|\frac{1}{K}\sum_{k=1}^{K}\nabla'F(\tilde{\mathbf{w}}_{\mathbf{c},k,t})\|^2, \qquad (5)
\end{aligned}
$$

where (a) follows from $\mathbf{w}_{\mathbf{c},k,t+1}=\tilde{\mathbf{w}}_{\mathbf{c},k,t}-\eta\tilde{\mathbf{g}}'_{\mathbf{c},k,t}$, (b) follows from

$$
\begin{aligned}
&\mathbb{E}\|\tilde{\mathbf{g}}'_{\mathbf{c},k,t}\|^2\\
&\overset{(a)}{=}\mathbb{E}\|\tilde{\mathbf{g}}'_{\mathbf{c},k,t}-\nabla'F(\tilde{\mathbf{w}}_{\mathbf{c},k,t})\|^2+\mathbb{E}\|\nabla'F(\tilde{\mathbf{w}}_{\mathbf{c},k,t})\|^2\\
&=\mathbb{E}\|Q(\tilde{\mathbf{g}}_{\mathbf{c},k,t}-\nabla F(\tilde{\mathbf{w}}_{\mathbf{c},k,t}))\|^2+\mathbb{E}\|\nabla'F(\tilde{\mathbf{w}}_{\mathbf{c},k,t})\|^2\\
&\overset{(b)}{=}\mathbb{E}\|\tilde{\mathbf{g}}_{\mathbf{c},k,t}-\nabla F(\tilde{\mathbf{w}}_{\mathbf{c},k,t})\|^2+\mathbb{E}\|\nabla F(\tilde{\mathbf{w}}_{\mathbf{c},k,t})\|^2\\
&=\mathbb{E}\|\mathbf{m}_{k,t}\odot(\mathbf{g}_{\mathbf{c},k,t}-\nabla F(\mathbf{w}_{\mathbf{c},k,t}))\|^2+\mathbb{E}\|\nabla F(\tilde{\mathbf{w}}_{\mathbf{c},k,t})\|^2\\
&\overset{(c)}{\le}\sum_{l=1}^{L_c}\sigma_l^2+\mathbb{E}\|\nabla'F(\tilde{\mathbf{w}}_{\mathbf{c},k,t})\|^2, \qquad (6)
\end{aligned}
$$

where (a) follows by the unbiased stochastic gradient Assumption 2 and the definition of variance, i.e., $\mathbb{E}[\|\mathbf{x}\|^2]=\mathbb{E}[\|\mathbf{x}-\mathbb{E}[\mathbf{x}]\|^2]+[\mathbb{E}\|\mathbf{x}\|]^2$; (b) and (c) follow from Assumption 6 and Assumption 3, respectively.

Similarly, $\mathbb{E}\|\mathbf{w}_{\mathbf{s},t+1}-\mathbf{w}_{\mathbf{s},t}\|^2$ can be bounded as

$$
\begin{aligned}
&\mathbb{E}\|\mathbf{w}_{\mathbf{s},t+1}-\mathbf{w}_{\mathbf{s},t}\|^2=\eta^2\mathbb{E}\|\frac{1}{K}\sum_{k=1}^{K}\mathbf{g}_{\mathbf{s},k,t}\|^2\\
&=\eta^2\mathbb{E}\|\frac{1}{K}\sum_{k=1}^{K}(\mathbf{g}_{\mathbf{s},k,t}-\nabla F(\mathbf{w}_{\mathbf{s},k,t}))\|^2\\
&+\eta^2\mathbb{E}\|\frac{1}{K}\sum_{k=1}^{K}\nabla F(\mathbf{w}_{\mathbf{s},k,t})\|^2\\
&\overset{(a)}{\le}\frac{\eta^2}{K}\sum_{l=L_c+1}^{L}\sigma_l^2+\eta^2\mathbb{E}\|\frac{1}{K}\sum_{k=1}^{K}\nabla F(\mathbf{w}_{\mathbf{s},k,t})\|^2, \qquad (7)
\end{aligned}
$$

where (a) follows from Assumption 3.

Thus, substituting Eqn. (5) and Eqn. (7) into Eqn. (4), $\mathbb{E}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|^2$ can be bounded as

$$
\begin{aligned}
\mathbb{E}\|\mathbf{w}_{t+1}-\mathbf{w}_t\|^2\le&\frac{\eta^2}{K}\sum_{l=1}^{L}\sigma_l^2+\frac{\eta^2}{K}\sum_{l=1}^{L_c}\sigma_l^2+\frac{2\rho_t}{K}\sum_{l=1}^{L_c}W_l^2\\
&+2\eta^2\mathbb{E}\|\frac{1}{K}\sum_{k=1}^{K}\nabla'F(\tilde{\mathbf{w}}_{\mathbf{c},k,t})\|^2+\eta^2\mathbb{E}\|\frac{1}{K}\sum_{k=1}^{K}\nabla F(\mathbf{w}_{\mathbf{s},k,t})\|^2 \qquad (8)
\end{aligned}
$$

We further note that

$$
\begin{aligned}
&\mathbb{E}\langle\nabla F(\mathbf{w}_t),\mathbf{w}_{t+1}-\mathbf{w}_t\rangle\\
=&\mathbb{E}\langle\nabla F(\mathbf{w}_{\mathbf{s},t}),\mathbf{w}_{\mathbf{s},t+1}-\mathbf{w}_{\mathbf{s},t}\rangle+\mathbb{E}\langle\nabla F(\mathbf{w}_{\mathbf{c},t}),\mathbf{w}_{\mathbf{c},t+1}-\mathbf{w}_{\mathbf{c},t}\rangle. \qquad (9)
\end{aligned}
$$

The first term can be written as

$$
\begin{aligned}
&\mathbb{E}\langle\nabla F(\mathbf{w}_{\mathbf{s},t}),\mathbf{w}_{\mathbf{s},t+1}-\mathbf{w}_{\mathbf{s},t}\rangle\\
=&-\eta\mathbb{E}\langle\nabla F(\mathbf{w}_{\mathbf{s},t}),\frac{1}{K}\sum_{k=1}^{K}\mathbf{g}_{\mathbf{s},k,t}\rangle\\
\overset{(a)}{=}&-\eta\mathbb{E}\langle\nabla F(\mathbf{w}_{\mathbf{c},t}),\frac{1}{K}\sum_{k=1}^{K}\nabla F(\mathbf{w}_{\mathbf{c},k,t})\rangle\\
\overset{(b)}{=}&-\frac{\eta}{2}\mathbb{E}\|\nabla F(\mathbf{w}_{\mathbf{s},t})\|^2-\frac{\eta}{2}\mathbb{E}\|\frac{1}{K}\sum_{k=1}^{K}\nabla F(\mathbf{w}_{\mathbf{s},k,t})\|^2\\
&+\frac{\eta}{2}\mathbb{E}\|\nabla F(\mathbf{w}_{\mathbf{s},t})-\frac{1}{K}\sum_{k=1}^{K}\nabla F(\mathbf{w}_{\mathbf{s},k,t})\|^2, \qquad (10)
\end{aligned}
$$

where (a) follows from Assumption 2; (b) follows from the identity $\langle\mathbf{a},\mathbf{b}\rangle=\frac{1}{2}(\|\mathbf{a}\|^2+\|\mathbf{b}\|^2-\|\mathbf{a}-\mathbf{b}\|^2)$. For the second term in Eqn. (9), we have

$$
\begin{aligned}
&\mathbb{E}\langle\nabla F(\mathbf{w}_{\mathbf{c},t}),\mathbf{w}_{\mathbf{c},t+1}-\mathbf{w}_{\mathbf{c},t}\rangle\\
\overset{(a)}{=}&\mathbb{E}\langle\nabla F(\mathbf{w}_{\mathbf{c},t}),\frac{1}{K}\sum_{k=1}^{K}(\tilde{\mathbf{w}}_{\mathbf{c},k,t}-\eta\tilde{\mathbf{g}}'_{\mathbf{c},k,t})-\mathbf{w}_{\mathbf{c},t}\rangle\\
=&\mathbb{E}\langle\nabla F(\mathbf{w}_{\mathbf{c},t}),-\eta\frac{1}{K}\sum_{k=1}^{K}\tilde{\mathbf{g}}'_{\mathbf{c},k,t}\rangle\\
&+\mathbb{E}\langle\nabla F(\mathbf{w}_{\mathbf{c},t}),\frac{1}{K}\sum_{k=1}^{K}\tilde{\mathbf{w}}_{\mathbf{c},k,t}-\mathbf{w}_{\mathbf{c},t}\rangle \qquad (11)
\end{aligned}
$$

where term $\mathbb{E}\langle\nabla F(\mathbf{w}_{\mathbf{c},t}),-\eta\frac{1}{K}\sum_{k=1}^{K}\tilde{\mathbf{g}}'_{\mathbf{c},k,t}\rangle$ in Eqn. (11) can be written as

$$
\begin{aligned}
&\mathbb{E}\langle\nabla F(\mathbf{w}_{\mathbf{c},t}),-\eta\frac{1}{K}\sum_{k=1}^{K}\tilde{\mathbf{g}}'_{\mathbf{c},k,t}\rangle\\
=&-\eta\mathbb{E}\langle\nabla F(\mathbf{w}_{\mathbf{c},t}),\frac{1}{K}\sum_{k=1}^{K}\tilde{\mathbf{g}}'_{\mathbf{c},k,t}\rangle\\
=&-\frac{\eta}{2}\mathbb{E}\|\nabla'F(\mathbf{w}_{\mathbf{c},t})\|^2-\frac{\eta}{2}\mathbb{E}\|\frac{1}{K}\sum_{k=1}^{K}\nabla'F(\tilde{\mathbf{w}}_{\mathbf{c},k,t})\|^2\\
&+\frac{\eta}{2}\mathbb{E}\|\nabla F(\mathbf{w}_{\mathbf{c},t})-\frac{1}{K}\sum_{k=1}^{K}\nabla'F(\tilde{\mathbf{w}}_{\mathbf{c},k,t})\|^2 \qquad (12)
\end{aligned}
$$

and term $\mathbb{E}\langle\nabla F(\mathbf{w}_{\mathbf{c},t}),\frac{1}{K}\sum_{k=1}^{K}\tilde{\mathbf{w}}_{\mathbf{c},k,t}-\mathbf{w}_{\mathbf{c},t}\rangle$ in Eqn. (11) can be bounded as

$$
\mathbb{E}\langle\nabla F(\mathbf{w}_{\mathbf{c},t}),\frac{1}{K}\sum_{k=1}^{K}\tilde{\mathbf{w}}_{\mathbf{c},k,t}-\mathbf{w}_{\mathbf{c},t}\rangle
$$

$$=\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\langle\nabla F(\mathbf{w}_{\mathbf{c},t}),\tilde{\mathbf{w}}_{\mathbf{c},k,t}-\mathbf{w}_{\mathbf{c},t}\rangle$$
$$\overset{(a)}{\leq}\frac{1}{2K}\sum_{k=1}^{K}(\mathbb{E}\|\nabla F(\mathbf{w}_{\mathbf{c},t})\|^2+\mathbb{E}\|\tilde{\mathbf{w}}_{\mathbf{c},k,t}-\mathbf{w}_{\mathbf{c},t}\|^2)$$
$$\overset{(b)}{\leq}\frac{1}{2}((8\eta^2(I+1)^2+1)\sum_{l=1}^{L_c}G_l^2+4\sum_{l=1}^{L}W_l^2+2\rho_t\sum_{l=1}^{L_c}W_l^2), \tag{13}$$

where (a) follows by the inequality $\langle\mathbf{a},\mathbf{b}\rangle\leq\frac{1}{2}(\|\mathbf{a}\|^2+\|\mathbf{b}\|^2)$; (b) follows from Assumption 4 and Lemma 2.

Substituting Eqn. (12) and Eqn. (13) into Eqn. (11), $\mathbb{E}\langle\nabla F(\mathbf{w}_{\mathbf{c},t}),\mathbf{w}_{\mathbf{c},t+1}-\mathbf{w}_{\mathbf{c},t}\rangle$ can be bounded as

$$\begin{aligned}
&\mathbb{E}\langle\nabla F(\mathbf{w}_{\mathbf{c},t}),\mathbf{w}_{\mathbf{c},t+1}-\mathbf{w}_{\mathbf{c},t}\rangle\\
\leq&-\frac{\eta}{2}\mathbb{E}\|\nabla F(\mathbf{w}_{\mathbf{c},t})\|^2-\frac{\eta}{2}\mathbb{E}\|\frac{1}{K}\sum_{k=1}^{K}\nabla' F(\tilde{\mathbf{w}}_{\mathbf{c},k,t})\|^2\\
&+\frac{\eta}{2}\mathbb{E}\|\nabla F(\mathbf{w}_{\mathbf{c},t})-\frac{1}{K}\sum_{k=1}^{K}\nabla' F(\tilde{\mathbf{w}}_{\mathbf{c},k,t})\|^2\\
&+\frac{1}{2}((8\eta^2(I+1)^2+1)\sum_{l=1}^{L_c}G_l^2+4\sum_{l=1}^{L}W_l^2+2\rho_t\sum_{l=1}^{L_c}W_l^2)
\end{aligned} \tag{14}$$

Substituting Eqn. (10) and Eqn. (14) into Eqn. (9), we have

$$\begin{aligned}
&\mathbb{E}\langle\nabla F(\mathbf{w}_t),\mathbf{w}_{t+1}-\mathbf{w}_t\rangle\\
\leq&-\frac{\eta}{2}\{\mathbb{E}\|\nabla F(\mathbf{w}_{\mathbf{s},t})\|^2+\mathbb{E}\|\nabla F(\mathbf{w}_{\mathbf{c},t})\|^2\}\\
&-\frac{\eta}{2}\{\mathbb{E}\|\frac{1}{K}\sum_{k=1}^{K}\nabla F(\mathbf{w}_{\mathbf{s},k,t})\|^2+\mathbb{E}\|\frac{1}{K}\sum_{k=1}^{K}\nabla' F(\tilde{\mathbf{w}}_{\mathbf{c},k,t})\|^2\}\\
&+\frac{\eta}{2}\{\mathbb{E}\|\nabla F(\mathbf{w}_{\mathbf{s},t})-\frac{1}{K}\sum_{k=1}^{K}\nabla F(\mathbf{w}_{\mathbf{s},k,t})\|^2\\
&\quad+\mathbb{E}\|\nabla F(\mathbf{w}_{\mathbf{c},t})-\frac{1}{K}\sum_{k=1}^{K}\nabla' F(\tilde{\mathbf{w}}_{\mathbf{c},k,t})\|^2\}\\
&+\frac{1}{2}((8\eta^2(I+1)^2+1)\sum_{l=1}^{L_c}G_l^2+4\sum_{l=1}^{L}W_l^2+2\rho_t\sum_{l=1}^{L_c}W_l^2)
\end{aligned} \tag{15}$$

Then substituting Eqn. (8) and Eqn. (15) into Eqn. (3), we have

$$\begin{aligned}
&\mathbb{E}[F(\mathbf{w}_{t+1})]\\
\leq&\mathbb{E}[F(\mathbf{w}_t)]-\frac{\eta}{2}\mathbb{E}\|\nabla F(\mathbf{w}_t)\|^2\\
&-\frac{\eta-\eta^2\beta}{2}\mathbb{E}\|\frac{1}{K}\sum_{k=1}^{K}\nabla F(\mathbf{w}_{\mathbf{s},k,t})\|^2\\
&-\frac{\eta-2\eta^2\beta}{2}\mathbb{E}\|\frac{1}{K}\sum_{k=1}^{K}\nabla' F(\tilde{\mathbf{w}}_{\mathbf{c},k,t})\|^2\\
&+\frac{\beta\eta^2}{2K}\sum_{l=1}^{L}\sigma_l^2+\frac{\beta\eta^2}{2K}\sum_{l=1}^{L_c}\sigma_l^2+\frac{\beta\rho_t}{K}\sum_{l=1}^{L_c}W_l^2\\
&+\frac{1}{2}((8\eta^2(I+1)^2+1)\sum_{l=1}^{L_c}G_l^2+4\sum_{l=1}^{L}W_l^2+2\rho_t\sum_{l=1}^{L_c}W_l^2)\\
&+\frac{\eta}{2}\{\mathbb{E}\|\nabla F(\mathbf{w}_{\mathbf{s},t})-\frac{1}{K}\sum_{k=1}^{K}\nabla F(\mathbf{w}_{\mathbf{s},k,t})\|^2\\
&+\mathbb{E}\|\nabla F(\mathbf{w}_{\mathbf{c},t})-\frac{1}{K}\sum_{k=1}^{K}\nabla' F(\tilde{\mathbf{w}}_{\mathbf{c},k,t})\|^2\}\\
\overset{(a)}{\leq}&\mathbb{E}[F(\mathbf{w}_t)]-\frac{\eta}{2}\mathbb{E}\|\nabla F(\mathbf{w}_t)\|^2\\
&+\frac{\beta\eta^2}{2K}\sum_{l=1}^{L}\sigma_l^2+\frac{\beta\eta^2}{2K}\sum_{l=1}^{L_c}\sigma_l^2+\frac{\beta\rho_t}{K}\sum_{l=1}^{L_c}W_l^2\\
&+\frac{1}{2}((8\eta^2(I+1)^2+1)\sum_{l=1}^{L_c}G_l^2+4\sum_{l=1}^{L}W_l^2+2\rho_t\sum_{l=1}^{L_c}W_l^2)\\
&+\frac{\eta}{2}\{\mathbb{E}\|\nabla F(\mathbf{w}_{\mathbf{s},t})-\frac{1}{K}\sum_{k=1}^{K}\nabla F(\mathbf{w}_{\mathbf{s},k,t})\|^2\\
&+\mathbb{E}\|\nabla F(\mathbf{w}_{\mathbf{c},t})-\frac{1}{K}\sum_{k=1}^{K}\nabla' F(\tilde{\mathbf{w}}_{\mathbf{c},k,t})\|^2\}\\
\overset{(b)}{\leq}&\mathbb{E}[F(\mathbf{w}_t)]-\frac{\eta}{2}\mathbb{E}\|\nabla F(\mathbf{w}_t)\|^2\\
&+\frac{\beta\eta^2}{2K}\sum_{l=1}^{L}\sigma_l^2+\frac{\beta\eta^2}{2K}\sum_{l=1}^{L_c}\sigma_l^2+\frac{\beta\rho_t}{K}\sum_{l=1}^{L_c}W_l^2+2\sum_{l=1}^{L_c}J_l^2\\
&+(2\beta^2+\frac{1}{2})((8\eta^2(I+1)^2+1)\sum_{l=1}^{L_c}G_l^2+4\sum_{l=1}^{L}W_l^2\\
&+2\rho_t\sum_{l=1}^{L_c}W_l^2)
\end{aligned} \tag{16}$$

where (a) follows from $0<\eta\leq\frac{1}{2\beta}$ and (b) holds because of the following inequality Eqn. (17) and Eqn. (18)

$$\begin{aligned}
&\mathbb{E}\|\nabla F(\mathbf{w}_{\mathbf{s},t})-\frac{1}{K}\sum_{k=1}^{K}\nabla F(\mathbf{w}_{\mathbf{s},k,t})\|^2\\
=&\mathbb{E}\|\frac{1}{K}\sum_{k=1}^{K}\nabla F(\mathbf{w}_{\mathbf{s},t})-\frac{1}{K}\sum_{k=1}^{K}\nabla F(\mathbf{w}_{\mathbf{s},k,t})\|^2\\
\leq&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\nabla F(\mathbf{w}_{\mathbf{s},t})-\nabla F(\mathbf{w}_{\mathbf{s},k,t})\|^2\\
\overset{(a)}{\leq}&\frac{\beta^2}{K}\sum_{k=1}^{K}\mathbb{E}\|\mathbf{w}_{\mathbf{s},t}-\mathbf{w}_{\mathbf{s},k,t}\|^2\overset{(b)}{=}0,
\end{aligned} \tag{17}$$

where (a) follows from Assumption 1; (b) holds because the server-side model of each client is the aggregated version of the whole server-side model. The term $\mathbb{E}\|\nabla F(\mathbf{w}_{\mathbf{c},t})-\frac{1}{K}\sum_{k=1}^{K}\nabla' F(\tilde{\mathbf{w}}_{\mathbf{c},k,t})\|^2$ in Eqn. (16) can be bounded as

$$\begin{aligned}
&\mathbb{E}\|\nabla F(\mathbf{w}_{\mathbf{c},t})-\frac{1}{K}\sum_{k=1}^{K}\nabla' F(\tilde{\mathbf{w}}_{\mathbf{c},k,t})\|^2\\
\leq&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\nabla F(\mathbf{w}_{\mathbf{c},t})-\nabla' F(\tilde{\mathbf{w}}_{\mathbf{c},k,t})\|^2\\
\leq&\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\|\nabla F(\mathbf{w}_{\mathbf{c},t})-\nabla F(\tilde{\mathbf{w}}_{\mathbf{c},k,t})
\end{aligned}$$

$$
\begin{aligned}
&+ \nabla F(\tilde{\mathbf{w}}_{\mathbf{c},k,t}) - \nabla' F(\tilde{\mathbf{w}}_{\mathbf{c},k,t})\|^2 \\
&\overset{(a)}{\leq} \frac{2}{K}\sum_{k=1}^{K}\mathbb{E}\{\|\nabla F(\mathbf{w}_{\mathbf{c},t}) - \nabla F(\tilde{\mathbf{w}}_{\mathbf{c},k,t})\|^2 + \sum_{l=1}^{L_c} J_l^2\} \\
&\overset{(b)}{\leq} \frac{2}{K}\sum_{k=1}^{K}\mathbb{E}\{\beta^2\|\mathbf{w}_{\mathbf{c},t} - \tilde{\mathbf{w}}_{\mathbf{c},k,t}\|^2 + \sum_{l=1}^{L_c} J_l^2\} \\
&\overset{(c)}{\leq} 2\beta^2((8\eta^2(I+1)^2+1)\sum_{l=1}^{L_c} G_l^2 + 4\sum_{l=1}^{L} W_l^2 \\
&+ 2\rho_t \sum_{l=1}^{L_c} W_l^2) + 2\sum_{l=1}^{L_c} J_l^2 \qquad (18)
\end{aligned}
$$

where (a) follows from the inequality $\|\sum_{i=1}^{n}\mathbf{z}_i\|^2 \leq n\sum_{i=1}^{n}\|\mathbf{z}_i\|^2$ and Assumption 6, (b) follows from Assumption 1 and (c) follows from Lemma 2.

Rearranging Eqn. (16) and dividing both sides by $\frac{\eta}{2T}$ and summing over $t \in \{1, ..., T\}$, the inequality can be written as

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F(\mathbf{w}_t)\|^2 \\
&\overset{(a)}{<} \frac{2(F(\mathbf{w}_1) - F(\mathbf{w}*))}{\eta T} \\
&+ \sum_{l=1}^{L}(\frac{\beta\eta}{K}\sigma_l^2 + \frac{1}{\eta}G_l^2 + \frac{4(4\beta^2+1)}{\eta}W_l^2) \\
&+ \sum_{l=1}^{L_c}(\frac{\beta\eta}{K}\sigma_l^2 + \frac{(4\beta^2+1)(8\eta^2(I+1)^2+1)}{\eta}G_l^2 \\
&+ \frac{\rho_f(4K\beta^2+K+\beta)}{K\eta}W_l^2 + \frac{4}{\eta}J_l^2) \\
&\overset{(b)}{\leq} \frac{2\vartheta}{\eta T} + \sum_{l=1}^{L}(\frac{\beta\eta}{K}\sigma_l^2 + \frac{1}{\eta}G_l^2 + \frac{4(4\beta^2+1)}{\eta}W_l^2) \\
&+ \sum_{l=1}^{L_c}(\frac{\beta\eta}{K}\sigma_l^2 + \frac{(4\beta^2+1)(8\eta^2(I+1)^2+1)}{\eta}G_l^2 \\
&+ \frac{\rho_f(4K\beta^2+K+\beta)}{K\eta}W_l^2 + \frac{4}{\eta}J_l^2), \qquad (19)
\end{aligned}
$$

where (a) follows from Lemma 1, (b) follows because $F(\mathbf{w}*)$ is the minimum value of problem.