Supplementary Material: "Enhancing Generalization in Federated Split Learning with Non-IID Data Toward Pervasive Intelligence"

Huiqing Ao, Hui Tian, *Senior Member, IEEE*, Wanli Ni, and Dusit Niyato, *Fellow, IEEE*

APPENDIX A
PROOF OF THEOREM 1

We first bound the convergence gap of the device-side models. We define $\bar{\boldsymbol{g}}_{n,t,\tau} = \nabla F_n(\mathbf{w}_{n,t,\tau})$, thus $\bar{\boldsymbol{g}}_{n,t,\tau} = \mathbb{E}[\boldsymbol{g}_{n,t,\tau}]$. For device n, we have

$$\begin{aligned}&\mathbb{E}[F_n(\mathbf{w}^{\mathrm{d}}_{n,t+1})] - \mathbb{E}[F_n(\mathbf{w}^{\mathrm{d}}_{n,t})]\\ &\leq \mathbb{E}\langle \nabla F_n(\mathbf{w}^{\mathrm{d}}_{n,t}), \mathbf{w}^{\mathrm{d}}_{n,t+1} - \mathbf{w}^{\mathrm{d}}_{n,t}\rangle + \frac{\ell}{2}\mathbb{E}[\|\mathbf{w}^{\mathrm{d}}_{n,t+1} - \mathbf{w}^{\mathrm{d}}_{n,t}\|^2].\end{aligned} \tag{1}$$

To bound the above inequality, we introduce the following lemmas.

Lemma 2. *According to Assumption 2, we can further obtain* $\mathbb{E}\|\nabla F_n(\mathbf{w}^{d}_{n,t,\tau}; \mathcal{B}_{n,t,\tau})\|^2 \leq lG^2$ *and* $\mathbb{E}\|\nabla F_n(\mathbf{w}^{s}_{n,t,\tau}; \mathcal{B}_{n,t,\tau})\|^2 \leq (L-l)G^2$, $\forall n, t, \tau$.

Lemma 3. *According to Assumption 3, we can further obtain* $\mathbb{E}\|\mathbf{g}^{d}_{n,t,\tau} - \mathbf{g}^{d}_{j,t,\tau}\|^2 \leq l\sigma^2_{n,j,t,\tau}, \forall n, j \in \mathcal{K}_{n,t,\tau}, t, \tau$.

Now, we bound the first term in (1) as follows:

$$\begin{aligned}&\mathbb{E}\langle \nabla F_n(\mathbf{w}^{\mathrm{d}}_{n,t}), \mathbf{w}^{\mathrm{d}}_{n,t+1} - \mathbf{w}^{\mathrm{d}}_{n,t}\rangle\\ &= -\sum_{\tau=0}^{H-1} \frac{\eta}{K_{n,t,\tau}+1}\mathbb{E}\langle \nabla F_n(\mathbf{w}^{\mathrm{d}}_{n,t}), \boldsymbol{g}^{\mathrm{d}}_{n,t,\tau} + \textstyle\sum_{j\in\mathcal{K}_{n,t,\tau}} \boldsymbol{g}^{\mathrm{d}}_{j,t,\tau}\rangle\\ &= -\sum_{\tau=0}^{H-1} \frac{\eta}{K_{n,t,\tau}+1}\mathbb{E}\langle \nabla F_n(\mathbf{w}^{\mathrm{d}}_{n,t}), \boldsymbol{g}^{\mathrm{d}}_{n,t,\tau}\rangle\\ &- \sum_{\tau=0}^{H-1} \frac{\eta}{K_{n,t,\tau}+1}\mathbb{E}\langle \nabla F_n(\mathbf{w}^{\mathrm{d}}_{n,t}), \textstyle\sum_{j\in\mathcal{K}_{n,t,\tau}} \boldsymbol{g}^{\mathrm{d}}_{j,t,\tau}\rangle.\end{aligned} \tag{2}$$

For the first term in (2), we have

$$\begin{aligned}&-\sum_{\tau=0}^{H-1} \frac{\eta}{K_{n,t,\tau}+1}\mathbb{E}\langle \nabla F_n(\mathbf{w}^{\mathrm{d}}_{n,t}), \boldsymbol{g}^{\mathrm{d}}_{n,t,\tau}\rangle\\ &\overset{(i)}{=} -\sum_{\tau=0}^{H-1} \frac{\eta}{2(K_{n,t,\tau}+1)}\mathbb{E}[\|\nabla F_n(\mathbf{w}^{\mathrm{d}}_{n,t})\|^2 + \|\nabla F_n(\mathbf{w}^{\mathrm{d}}_{n,t,\tau})\|^2\\ &- \|\nabla F_n(\mathbf{w}^{\mathrm{d}}_{n,t}) - \nabla F_n(\mathbf{w}^{\mathrm{d}}_{n,t,\tau})\|^2]\\ &\overset{(ii)}{\leq} -\sum_{h=0}^{H-1} \frac{\eta}{2(K_{n,t,\tau}+1)}\mathbb{E}[\|\nabla F_n(\mathbf{w}^{\mathrm{d}}_{n,t})\|^2]\\ &+ \sum_{\tau=0}^{H-1} \frac{\eta\ell^2}{2(K_{n,t,\tau}+1)}\mathbb{E}[\|\mathbf{w}^{\mathrm{d}}_{n,t} - \mathbf{w}^{\mathrm{d}}_{n,t,\tau}\|^2]\\ &= -\sum_{\tau=0}^{H-1} \frac{\eta}{2(K_{n,t,\tau}+1)}\mathbb{E}[\|\nabla F_n(\mathbf{w}^{\mathrm{d}}_{n,t})\|^2]\\ &+ \sum_{\tau=0}^{H-1} \frac{\eta\ell^2}{2(K_{n,t,\tau}+1)}\mathbb{E}[\|\sum_{m=0}^{\tau-1}\nabla F_n(\mathbf{w}^{\mathrm{d}}_{n,t,m})\|^2]\\ &\overset{(iii)}{\leq} -\sum_{\tau=0}^{H-1} \frac{\eta}{2(K_{n,t,\tau}+1)}\mathbb{E}[\|\nabla F_n(\mathbf{w}^{\mathrm{d}}_{n,t})\|^2]\\ &+ \sum_{\tau=0}^{H-1} \frac{\eta\ell^2}{2(K_{n,t,\tau}+1)}\tau\sum_{m=0}^{\tau-1}\mathbb{E}[\|\nabla F_n(\mathbf{w}^{\mathrm{d}}_{n,t,m})\|^2]\\ &\overset{(iv)}{\leq} -\sum_{\tau=0}^{H-1} \frac{\eta\mathbb{E}[\|\nabla F_n(\mathbf{w}^{\mathrm{d}}_{n,t})\|^2]}{2(K_{n,t,\tau}+1)} + \sum_{\tau=0}^{H-1}\frac{\eta l\ell^2\tau^2G^2}{2(K_{n,t,\tau}+1)},\end{aligned} \tag{3}$$

where (i) follows by using $\langle \boldsymbol{a}, \boldsymbol{b}\rangle = \frac{1}{2}\|\boldsymbol{a}\|^2 + \frac{1}{2}\|\boldsymbol{b}\|^2 - \frac{1}{2}\|\boldsymbol{a} - \boldsymbol{b}\|^2$, (ii) follows Assumption 1, (iii) follows by using $\|\sum_{m=0}^{\tau-1}\boldsymbol{a}_m\|^2 \leq \tau\sum_{m=0}^{\tau-1}\|\boldsymbol{a}_m\|^2$, and (iv) follows Assumption 2.

For the second term in (2), we have

$$\begin{aligned}&-\sum_{\tau=0}^{H-1} \frac{\eta K_{n,t,\tau}}{K_{n,t,\tau}+1}\mathbb{E}\langle \nabla F_n(\mathbf{w}^{\mathrm{d}}_{n,t}), \frac{1}{K_{n,t,\tau}}\textstyle\sum_{j\in\mathcal{K}_{n,t,\tau}} \boldsymbol{g}^{\mathrm{d}}_{j,t,\tau}\rangle\\ &\overset{(i)}{=} -\sum_{\tau=0}^{H-1} \frac{\eta K_{n,t,\tau}}{K_{n,t,\tau}+1}\frac{1}{2}\mathbb{E}[\|\nabla F_n(\mathbf{w}^{\mathrm{d}}_{n,t})\|^2\\ &+ \|\frac{1}{K_{n,t,\tau}}\textstyle\sum_{j\in\mathcal{K}_{n,t,\tau}}\nabla F_j(\mathbf{w}^{\mathrm{d}}_{j,t,\tau})\|^2\\ &- \|\nabla F_n(\mathbf{w}^{\mathrm{d}}_{n,t}) - \frac{1}{K_{n,t,\tau}}\textstyle\sum_{j\in\mathcal{K}_{n,t,\tau}}\nabla F_j(\mathbf{w}^{\mathrm{d}}_{j,t,\tau})\|^2]\\ &\overset{(ii)}{\leq} -\sum_{\tau=0}^{H-1} \frac{\eta K_{n,t,\tau}}{2(K_{n,t,\tau}+1)}\mathbb{E}[\|\nabla F_n(\mathbf{w}^{\mathrm{d}}_{n,t})\|^2\\ &+ \sum_{\tau=0}^{H-1} \frac{\eta K_{n,t,\tau}}{K_{n,t,\tau}+1}\mathbb{E}[\|\nabla F_n(\mathbf{w}^{\mathrm{d}}_{n,t}) - \nabla F_n(\mathbf{w}^{\mathrm{d}}_{n,t,\tau})\|^2\\ &+ \|\nabla F_n(\mathbf{w}^{\mathrm{d}}_{n,t,\tau}) - \frac{1}{K_{n,t,\tau}}\textstyle\sum_{j\in\mathcal{K}_{n,t,\tau}}\nabla F_j(\mathbf{w}^{\mathrm{d}}_{j,t,\tau})\|^2]\\ &\overset{(iii)}{\leq} -\sum_{\tau=0}^{H-1} \frac{\eta K_{n,t,\tau}}{2(K_{n,t,\tau}+1)}\mathbb{E}[\|\nabla F_n(\mathbf{w}^{\mathrm{d}}_{n,t})\|^2\end{aligned}$$

H. Ao and H. Tian 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).

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).

$$+\sum_{\tau=0}^{H-1}\frac{\eta l\ell^2\tau^2G^2K_{n,t,\tau}}{K_{n,t,\tau}+1}+\sum_{\tau=0}^{H-1}\frac{\eta\sum_{j\in\mathcal{K}_{n,t,\tau}}l\sigma_{n,j,t,\tau}^2}{K_{n,t,\tau}+1}, \tag{4}$$

where (i) follows by using $\langle \boldsymbol{a},\boldsymbol{b}\rangle=\frac{1}{2}\|\boldsymbol{a}\|^2+\frac{1}{2}\|\boldsymbol{b}\|^2-\frac{1}{2}\|\boldsymbol{a}-\boldsymbol{b}\|^2$, (ii) follows by using $\|\boldsymbol{a}+\boldsymbol{c}+\boldsymbol{b}+\boldsymbol{d}\|^2\leq 2\|\boldsymbol{a}+\boldsymbol{c}\|^2+2\|\boldsymbol{b}+\boldsymbol{d}\|^2$, (iii) follows by using $\|\sum_{m=0}^{\tau-1}\boldsymbol{a}_m\|^2\leq\tau\sum_{m=0}^{\tau-1}\|\boldsymbol{a}_m\|^2$, Assumptions 1, 2 and 3.

Plugging (3) and (4) into (2), we obtain

$$\begin{aligned}
&\mathbb{E}\langle\nabla F_n(\mathbf{w}_{n,t}^{\mathrm{d}}),\mathbf{w}_{n,t+1}^{\mathrm{d}}-\mathbf{w}_{n,t}^{\mathrm{d}}\rangle\\
&\leq-\frac{\eta H}{2}\mathbb{E}[\|\nabla F_n(\mathbf{w}_{n,t}^{\mathrm{d}})\|^2]+\sum_{\tau=0}^{H-1}\frac{\eta l\ell^2\tau^2G^2(2K_{n,t,\tau}+1)}{2(K_{n,t,\tau}+1)}\\
&+\sum_{\tau=0}^{H-1}\frac{\eta\sum_{j\in\mathcal{K}_{n,t,\tau}}l\sigma_{n,j,t,\tau}^2}{K_{n,t,\tau}+1}\\
&\leq-\frac{\eta H}{2}\mathbb{E}[\|\nabla F_n(\mathbf{w}_{n,t}^{\mathrm{d}})\|^2]+\sum_{\tau=0}^{H-1}\eta lH\ell^2\tau^2G^2\\
&+\sum_{\tau=0}^{H-1}\frac{\eta\sum_{j\in\mathcal{K}_{n,t,\tau}}l\sigma_{n,j,t,\tau}^2}{K_{n,t,\tau}+1}\\
&\overset{(i)}{\leq}-\frac{\eta H}{2}\mathbb{E}[\|\nabla F_n(\mathbf{w}_{n,t}^{\mathrm{d}})\|^2]+\frac{\eta lH(H-1)(2H-1)\ell^2G^2}{6}\\
&+\sum_{\tau=0}^{H-1}\frac{\eta\sum_{j\in\mathcal{K}_{n,t,\tau}}l\sigma_{n,j,t,\tau}^2}{K_{n,t,\tau}+1},
\end{aligned} \tag{5}$$

where (i) follows by using $\sum_{\tau=0}^{H-1}\tau^2=\frac{H(H-1)(2H-1)}{6}$.

Next, we bound the second term in (1) as follows:

$$\begin{aligned}
&\frac{\ell}{2}\mathbb{E}[\|\mathbf{w}_{n,t+1}^{\mathrm{d}}-\mathbf{w}_{n,t}^{\mathrm{d}}\|^2]\\
&=\frac{\ell}{2}\mathbb{E}[\|-\sum_{\tau=0}^{H-1}\frac{\eta}{K_{n,t,\tau}+1}(\boldsymbol{g}_{n,t,\tau}^{\mathrm{d}}+\sum\nolimits_{j\in\mathcal{K}_{n,t,\tau}}\boldsymbol{g}_{j,t,\tau}^{\mathrm{d}})\|^2]\\
&\overset{(i)}{\leq}\frac{\ell}{2}H\sum_{\tau=0}^{H-1}\frac{\eta^2}{(K_{n,t,\tau}+1)^2}\mathbb{E}[\|\boldsymbol{g}_{n,t,\tau}^{\mathrm{d}}+\sum\nolimits_{j\in\mathcal{K}_{n,t,\tau}}\boldsymbol{g}_{j,t,\tau}^{\mathrm{d}}\|^2]\\
&\overset{(ii)}{\leq}H\sum_{\tau=0}^{H-1}\frac{\ell\eta^2}{(K_{n,t,\tau}+1)^2}\mathbb{E}[\|\boldsymbol{g}_{n,t,\tau}^{\mathrm{d}}\|^2+\|\sum\nolimits_{j\in\mathcal{K}_{n,t,\tau}}\boldsymbol{g}_{j,t,\tau}^{\mathrm{d}}\|^2]\\
&\overset{(iii)}{\leq}H\sum_{\tau=0}^{H-1}\frac{\ell\eta^2\mathbb{E}[\|\boldsymbol{g}_{n,t,\tau}^{\mathrm{d}}\|^2+K_{n,t,\tau}\sum_{j\in\mathcal{K}_{n,t,\tau}}\|\boldsymbol{g}_{j,t,\tau}^{\mathrm{d}}\|^2]}{(K_{n,t,\tau}+1)^2}\\
&\overset{(iv)}{\leq}H\sum_{\tau=0}^{H-1}\frac{\ell\eta^2(lG^2(1+(K_{n,t,\tau})^2))}{(K_{n,t,\tau}+1)^2}\\
&\leq\ell l\eta^2H^2G^2,
\end{aligned} \tag{6}$$

where (i) and (iii) follow by using $\|\sum_{j\in\mathcal{K}_{n,t,\tau}}\boldsymbol{a}_j\|^2\leq K_{n,t,\tau}\sum_{j\in\mathcal{K}_{n,t,\tau}}\|\boldsymbol{a}_j\|^2$, (ii) follows by using $\|\boldsymbol{a}+\boldsymbol{b}\|^2\leq 2\|\boldsymbol{a}\|^2+2\|\boldsymbol{b}\|^2$, and (iv) follows Assumptions 2.

Plugging (6) and (5) into (1), we obtain

$$\begin{aligned}
&\mathbb{E}[F_n(\mathbf{w}_{n,t+1}^{\mathrm{d}})]-\mathbb{E}[F_n(\mathbf{w}_{n,t}^{\mathrm{d}})]\\
&\leq-\frac{\eta H}{2}\mathbb{E}[\|\nabla F_n(\mathbf{w}_{n,t}^{\mathrm{d}})\|^2]+\frac{\eta lH(H-1)(2H-1)\ell^2G^2}{6}\\
&+\sum_{\tau=0}^{H-1}\frac{\eta\sum_{j\in\mathcal{K}_{n,t,\tau}}l\sigma_{n,j,t,\tau}^2}{K_{n,t,\tau}+1}+\ell l\eta^2H^2G^2.
\end{aligned} \tag{7}$$

Recursively utilizing the above inequality from $t=0$ to $t=T-1$, we obtain

$$\begin{aligned}
&\mathbb{E}[F_n(\mathbf{w}_{n,T}^{\mathrm{d}})]-\mathbb{E}[F_n(\mathbf{w}_{n,0}^{\mathrm{d}})]\\
&\leq-\sum_{t=0}^{T-1}\frac{\eta H}{2}\mathbb{E}[\|\nabla F_n(\mathbf{w}_{n,t}^{\mathrm{d}})\|^2]+\frac{\eta lTH(H-1)(2H-1)\ell^2G^2}{6}\\
&+\sum_{t=0}^{T-1}\sum_{\tau=0}^{H-1}\frac{\eta\sum_{j\in\mathcal{K}_{n,t,\tau}}l\sigma_{n,j,t,\tau}^2}{K_{n,t,\tau}+1}+T\ell l\eta^2H^2G^2.
\end{aligned} \tag{8}$$

Dividing T for both sides of the above inequality and rearranging the terms, we can obtain

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla F_n(\mathbf{w}_{n,t}^{\mathrm{d}})\|^2]\\
&\leq\frac{2}{\eta HT}(\mathbb{E}[F_n(\mathbf{w}_0^{\mathrm{d}})]-\mathbb{E}[F_n(\mathbf{w}_{n,T}^{\mathrm{d}})])+\frac{l(H-1)(2H-1)\ell^2G^2}{3}\\
&+\frac{2}{HT}\sum_{t=0}^{T-1}\sum_{\tau=0}^{H-1}\frac{\sum_{j\in\mathcal{K}_{n,t,\tau}}l\sigma_{n,j,t,\tau}^2}{K_{n,t,\tau}+1}+2\ell l\eta HG^2,
\end{aligned} \tag{9}$$

where $\mathbf{w}_{n,0}^{\mathrm{d}}=\mathbf{w}_0^{\mathrm{d}}$, and $\eta>0$.

Since there are N different device-side models, we analyse the expected average convergence as follows:

$$\begin{aligned}
&\frac{1}{NT}\sum_{t=0}^{T-1}\sum_{n=1}^{N}\mathbb{E}[\|\nabla F_n(\mathbf{w}_{n,t}^{\mathrm{d}})\|^2]\leq\frac{2}{\eta HT}(\mathbb{E}[F(\mathbf{w}_0^{\mathrm{d}})]-\\
&\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[F_n(\mathbf{w}_{n,T}^{\mathrm{d}})])+\frac{(\ell(H-1)(2H-1)+6\eta H)\ell lG^2}{3}\\
&+\frac{2}{NHT}\sum_{t=0}^{T-1}\sum_{n=1}^{N}\sum_{\tau=0}^{H-1}\frac{\sum_{j\in\mathcal{K}_{n,t,\tau}}l\sigma_{n,j,t,\tau}^2}{K_{n,t,\tau}+1},
\end{aligned} \tag{10}$$

Furthermore, according to Jensen's inequality, we have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\frac{1}{N}\sum_{n=1}^{N}\nabla F_n(\mathbf{w}_{n,t}^{\mathrm{d}})\|^2]\leq\frac{2}{\eta HT}(\mathbb{E}[F(\mathbf{w}_0^{\mathrm{d}})]-\\
&\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[F_n(\mathbf{w}_{n,T}^{\mathrm{d}})])+\frac{(\ell(H-1)(2H-1)+6\eta H)\ell lG^2}{3}\\
&+\frac{2}{NHT}\sum_{t=0}^{T-1}\sum_{n=1}^{N}\sum_{\tau=0}^{H-1}\frac{\sum_{j\in\mathcal{K}_{n,t,\tau}}l\sigma_{n,j,t,\tau}^2}{K_{n,t,\tau}+1}.
\end{aligned} \tag{11}$$

Next, we bound the convergence gap of the server-side global model. Given Assumption 1, the server-side global loss function $F(\mathbf{w}^{\mathrm{s}})$ is ℓ-smooth function. Hence, we have

$$\begin{aligned}
&\mathbb{E}[F(\mathbf{w}_{t+1}^{\mathrm{s}})]-\mathbb{E}[F(\mathbf{w}_t^{\mathrm{s}})]\\
&\leq\mathbb{E}\langle\nabla F(\mathbf{w}_t^{\mathrm{s}}),\mathbf{w}_{t+1}^{\mathrm{s}}-\mathbf{w}_t^{\mathrm{s}}\rangle+\frac{\ell}{2}\mathbb{E}[\|\mathbf{w}_{t+1}^{\mathrm{s}}-\mathbf{w}_t^{\mathrm{s}}\|^2]=\\
&\mathbb{E}\langle\nabla F(\mathbf{w}_t^{\mathrm{s}}),-\sum_{n=1}^{N}\sum_{\tau=0}^{H-1}\frac{\eta}{N(K_{n,t,\tau}+1)}(\boldsymbol{g}_{n,t,\tau}^{\mathrm{s}}+\sum\nolimits_{j\in\mathcal{K}_{n,t,\tau}}\boldsymbol{g}_{j,t,\tau}^{\mathrm{s}})\rangle\\
&+\frac{\ell}{2}\mathbb{E}[\|-\sum_{n=1}^{N}\sum_{\tau=0}^{H-1}\frac{\eta}{N(K_{n,t,\tau}+1)}(\boldsymbol{g}_{n,t,\tau}^{\mathrm{s}}+\sum\nolimits_{j\in\mathcal{K}_{n,t,\tau}}\boldsymbol{g}_{j,t,\tau}^{\mathrm{s}})\|^2].
\end{aligned} \tag{12}$$

Now, we bound the first term in (12) as follows:

$$\mathbb{E}\langle\nabla F(\mathbf{w}_t^{\mathrm{s}}), -\sum_{n=1}^{N}\sum_{\tau=0}^{H-1}\frac{\eta}{N(K_{n,t,\tau}+1)}(\boldsymbol{g}_{n,t,\tau}^{\mathrm{s}}+\sum\nolimits_{j\in\mathcal{K}_{n,t,\tau}}\boldsymbol{g}_{j,t,\tau}^{\mathrm{s}})\rangle \stackrel{(i)}{\le} -\eta H\delta\mathbb{E}\|\nabla F(\mathbf{w}_t^{\mathrm{s}})\|^2, \quad (13)$$

where (i) follows (6).

Next, we bound the second term in (12) as follows:

$$\begin{aligned}&\frac{\ell}{2}\mathbb{E}[\|-\sum_{n=1}^{N}\sum_{\tau=0}^{H-1}\frac{\eta}{N(K_{n,t,\tau}+1)}(\boldsymbol{g}_{n,t,\tau}^{\mathrm{s}}+\sum\nolimits_{j\in\mathcal{K}_{n,t,\tau}}\boldsymbol{g}_{j,t,\tau}^{\mathrm{s}})\|^2]\\
&\stackrel{(i)}{\le}\frac{\ell H\eta^2}{2N(K_{n,t,\tau}+1)^2}\sum_{n=1}^{N}\sum_{\tau=0}^{H-1}\mathbb{E}[\|\boldsymbol{g}_{n,t,\tau}^{\mathrm{s}}+\sum\nolimits_{j\in\mathcal{K}_{n,t,\tau}}\boldsymbol{g}_{j,t,\tau}^{\mathrm{s}}\|^2]\\
&\stackrel{(ii)}{\le}\frac{\ell H\eta^2}{N(K_{n,t,\tau}+1)^2}\sum_{n=1}^{N}\sum_{\tau=0}^{H-1}\mathbb{E}[\|\boldsymbol{g}_{n,t,\tau}^{\mathrm{s}}\|^2+\|\sum\nolimits_{j\in\mathcal{K}_{n,t,\tau}}\boldsymbol{g}_{j,t,\tau}^{\mathrm{s}}\|^2]\\
&\stackrel{(iii)}{\le}\sum_{n=1}^{N}\sum_{\tau=0}^{H-1}\frac{\ell H(L-l)\eta^2G^2(1+(K_{n,t,\tau})^2)}{N(K_{n,t,\tau}+1)^2}\\
&\le \ell H^2(L-l)\eta^2G^2, \quad (14)\end{aligned}$$

where (i) follows by using Jensen's inequality, (ii) follows by using $\|\boldsymbol{a}+\boldsymbol{b}\|^2\le 2\|\boldsymbol{a}\|^2+2\|\boldsymbol{b}\|^2$, and (iii) follows Assumption 2 and $\|\sum_{j\in\mathcal{K}_{n,t,\tau}}\boldsymbol{a}_j\|^2\le K_{n,t,\tau}\sum_{j\in\mathcal{K}_{n,t,\tau}}\|\boldsymbol{a}_j\|^2$.

Plugging (14) and (13) into (12), we can obtain

$$\mathbb{E}[F(\mathbf{w}_{t+1}^{\mathrm{s}})]-\mathbb{E}[F(\mathbf{w}_t^{\mathrm{s}})]\le -\eta H\delta\mathbb{E}\|\nabla F(\mathbf{w}_t^{\mathrm{s}})\|^2+\ell H^2(L-l)\eta^2G^2. \quad (15)$$

Recursively utilizing the above inequality from $t=0$ to $t=T-1$ and then averaging yields

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(\mathbf{w}_t^{\mathrm{s}})\|^2\le\frac{\mathbb{E}[F(\mathbf{w}_0^{\mathrm{s}})]-\mathbb{E}[F(\mathbf{w}_T^{\mathrm{s}})]}{\eta TH\delta}+\frac{1}{\delta}\ell H(L-l)\eta G^2, \quad (16)$$

where $\eta>0$.

Now, we complete the proof of Theorem 1.

Appendix B

It is obvious that problem (28) is a convex problem if constraints (27c) and (27d) are convex, as other constraints and the objection are linear sets.

To prove this, we first define a function as

$$\varphi(x)=x\log_2(1+\frac{1}{x}), x>0. \quad (17)$$

Then, we can deduce that

$$\varphi^{'}(x)=\frac{1}{\ln(2)}(\ln(1+\frac{1}{x})-\frac{1}{x+1}), x>0 \quad (18)$$

and

$$\varphi^{''}(x)=-\frac{1}{\ln(2)x(1+x)^2}<0, x>0. \quad (19)$$

Clearly, $\varphi^{'}(x)$ is a decreasing function as $\varphi^{''}(x)<0$. Additionally, since $\lim_{\varphi^{'}(x)\to+\infty}=0$, we have $\varphi^{'}(x)>0$ for all $0<x<+\infty$. Thus, $\varphi(x)$ is an increasing function and $\varphi(x)>0$ with $x>0$. Furthermore, we can deduce that

$$(\frac{1}{\varphi(x)})^{'}=-\frac{\varphi^{'}(x)}{(\varphi(x))^2}<0, x>0 \quad (20)$$

and

$$(\frac{1}{\varphi(x)})^{''}=-\frac{\varphi^{''}(x)(\varphi(x))^2-2(\varphi^{'}(x))^2}{(\varphi(x))^4}>0, x>0. \quad (21)$$

Thus, $\frac{1}{\varphi(x)}$ is a convex function. Moreover, we obtain that the right hand sides of (27c) and (27d) are convex functions of bandwidths. Since there is no couple among variables, constraints (27c) and (27d) are convex. Therefore, problem (28) is a convex and can be effectively solved with some available toolkits such as CVXPY in python.