1 FedCET with Nonconvex Loss functions

We provide the convergence of FedCET under nonconvex loss functions with deterministic and stochastic cases.

Theorem 1. *(Deterministic Case) Under Assumption 1, there exists stepsize α and the parameter $0 < c < \frac{2}{(\tau+3)\alpha}$ such that*

$$\frac{1}{\tau K}\sum_{t=1}^{\tau K}\|\nabla f(\bar{x}(t))\|^2 \le O(\frac{1}{\tau K}).$$

where $\bar{x}(t) = \frac{1}{N}\sum_{i=1}^{N} x_i(t)$ and K is the communication round of FedCET.

Proof. See Section 2. □

Theorem 2. *(Stochastic Case) Under Assumption 1, there exists stepsize α and the parameter $0 < c < \frac{2}{(\tau+3)\alpha}$ such that*

$$\frac{1}{\tau K}\sum_{t=1}^{\tau K}\mathbb{E}\Big[\|\nabla f(\bar{x}(t))\|^2\Big] \le O(\frac{1}{\tau K}) + O(\tau\alpha\sigma^2).$$

where $\bar{x}(t) = \frac{1}{N}\sum_{i=1}^{N} x_i(t)$ and K is the communication round of FedCET.

Proof. See Section 3. □

2 Proof of Theorem 1

To analyze the convergence of FedCET, we define:

$$\begin{aligned}
\xi =& c\alpha,\\
X(t) =& [x_1(t), x_2(t), \cdots, x_N(t)],\\
\nabla f(X(t)) =& [\nabla f_1(x_1(t)), \nabla f_2(x_2(t)), \cdots, \nabla f_N(x_N(t))],\\
\overline{X}(t) =& \frac{1}{N}\sum_{i=1}^{N} x_i(t),\\
\overline{\nabla f}(X(t)) =& \frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i(t)),\\
W =& \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^{\mathbf{T}}.
\end{aligned}$$

In addition, we define the time-varying mixing matrix

$$W(t) = \begin{cases}(1-\xi)\mathbf{I}_N + \xi W, & t = k\tau,\\ \mathbf{I}_N, & t \neq k\tau\end{cases} \tag{1}$$

to represent the local updates of FedCET equivalently. Thus, (2) and (3) can be equivalently represented as

$$X(t+1) = 2X(t)W(t) - X(t-1)W(t) - \alpha\nabla f(X(t))W(t) + \alpha\nabla f(X(t-1))W(t). \tag{2}$$

From (2), we have

$$\overline{X}(t+1) = 2\overline{X}(t) - \overline{X}(t-1) - \alpha\overline{\nabla f}(X(t)) + \alpha\overline{\nabla f}(X(t-1)). \tag{3}$$

From (1) and (3), we have

$$\begin{aligned}
&\overline{X}(t+1) - \overline{X}(t)\\
=&\overline{X}(t) - \overline{X}(t-1) - \alpha\{\overline{\nabla f}(X(t)) - \overline{\nabla f}(X(t-1))\}\\
=&\overline{X}(0) - \overline{X}(-1) - \alpha\{\overline{\nabla f}(X(t)) - \overline{\nabla f}(X(-1))\}.
\end{aligned} \tag{4}$$

For convenience, we consider the initial value setting of $x_i(-1)$ and $x_i(0)$, which should follow the rules: $x_i(-1)$ can be arbitrarily chosen in $\mathbb{R}^n$ whereas $x_i(0)$ should be set as

$$x_i(0) = x_i(-1) - \alpha \nabla f_i(x_i(-1)). \tag{5}$$

From (4) and the initial value setting (5), we can obtain

$$\overline{X}(t+1) = \overline{X}(t) - \alpha \overline{\nabla f}(X(t)). \tag{6}$$

From (6), Assumption 1, and the property $\|a+b\|^2 = \|a\|^2 + \|b\|^2 + 2\langle a, b\rangle$, we have

$$\begin{aligned}
&f(\overline{X}(t+1)) \\
\leq& f(\overline{X}(t)) - \alpha \langle \nabla f(\overline{X}(t)), \overline{\nabla f}(X(t))\rangle + \frac{L\alpha^2}{2}\|\overline{\nabla f}(X(t))\|^2 \\
=& f(\overline{X}(t)) - \frac{\alpha}{2}\|\nabla f(\overline{X}(t))\|^2 - (\frac{\alpha}{2} - \frac{L\alpha^2}{2})\|\overline{\nabla f}(X(t))\|^2 + \frac{\alpha}{2}\|\nabla f(\overline{X}(t)) - \overline{\nabla f}(X(t))\|^2 \\
\leq& f(\overline{X}(t)) - \frac{\alpha}{2}\|\nabla f(\overline{X}(t))\|^2 - (\frac{\alpha}{2} - \frac{L\alpha^2}{2})\|\overline{\nabla f}(X(t))\|^2 + \frac{\alpha L^2}{2N}\sum_{i=1}^{N}\|\overline{X}(t) - x_i(t)\|^2.
\end{aligned} \tag{7}$$

Thus, from (7), we have

$$\begin{aligned}
&\sum_{t=1}^{T}\{\|\nabla f(\overline{X}(t))\|^2 + (1 - L\alpha)\|\overline{\nabla f}(X(t))\|^2\} \\
\leq& \frac{2}{\alpha}\{f(\overline{X}(1)) - f(x^*)\} + \frac{L^2}{N}\sum_{t=1}^{T}\sum_{i=1}^{N}\|\overline{X}(t) - x_i(t)\|^2.
\end{aligned} \tag{8}$$

From (8), it is necessary to analyze the consensus error $\sum_{t=1}^{T}\sum_{i=1}^{N}\|\overline{X}(t) - x_i(t)\|^2$ before establishing the convergence rate of FedCET. We define $\widetilde{W} = (1-\xi)\mathbf{I}_N + \xi W$ and thus eigenvalues $\{\rho_1, \rho_2, \cdots, \rho_N\}$ of $\widetilde{W}$ satisfies

$$\rho_i = 1 - \xi + \xi\lambda_i,$$

where

$$\lambda_i = \begin{cases} 1, & i = 1, \\ 0, & 2 \leq i \leq N, \end{cases} \tag{9}$$

are eigenvalues of W. From the definition $\xi = c\alpha$ and $0 < c < \frac{2}{(\tau+3)\alpha}$, we have

$$\rho_1 = 1, \quad \frac{\tau - 1}{\tau + 3} < \rho_i < 1, \quad \text{for } i = 2, 3, \cdots, N. \tag{10}$$

From (1), $W(t)$ can be equivalently represented via an orthogonal matrix $P = [v_1, v_2, \cdots, v_N] \in \mathbb{R}^{N\times N}$ satisfying $P^{\mathbf{T}}P = PP^{\mathbf{T}} = \mathbf{I}_N$ and a diagonal matrix $\Lambda(t) = \mathrm{diag}\{\rho_1(t), \rho_2(t), \cdots, \rho_N(t)\}$ as follows:

$$W(t) = P\Lambda(t)P^{\mathbf{T}}, \tag{11}$$

where $\rho_i(t) = \rho_i$ for $t = k\tau$ and $\rho_i(t) = 1$ for $t \neq k\tau$. Moreover, if we define

$$\begin{cases} Y(t) = X(t)P = [y_1(t), y_2(t), \cdots, y_N(t)], \\ H(t) = \overline{\nabla f}(X(t))P = [h_1(t), h_2(t), \cdots, h_N(t)], \end{cases} \tag{12}$$

from (2) and (11), we have

$$Y(t+1) = 2Y(t)\Lambda(t) - Y(t-1)\Lambda(t) - \alpha H(t)\Lambda(t) + \alpha H(t-1)\Lambda(t). \tag{13}$$

Moreover, from (12) and (13), we have

$$y_i(t+1) = \rho_i(t)(2y_i(t) - y_i(t-1) - \alpha h_i(t) + \alpha h_i(t-1)), \tag{14}$$

for $i = 1, 2, \cdots, N$. We denote the n-dimensional unit vector set as $\{e_i\}_{i=1,2,\cdots,n}$, where $[e_i]_j = 1$ for $j = i$ and $[e_i]_j = 0$ for $j \neq i$. Then, we present the relationship between $y_i(t)$ and $\overline{X}(t) - x_i(t)$ as follows:

$$
\begin{aligned}
&\sum_{i=1}^{N} \|\overline{X}(t) - x_i(t)\|^2 \\
=&\sum_{i=1}^{N} \|X(t)e_i - \frac{1}{N} X(t)\mathbf{1}_N\|^2 \\
=&\|X(t)v_1 v_1^{\mathbf{T}} - X(t)PP^{\mathbf{T}}\|_F^2 \\
=&\|X(t)P(\mathbf{I}_N - [e_1, \mathbf{0}_N, \cdots, \mathbf{0}_N])P^{\mathbf{T}}\|_F^2 \\
=&\sum_{i=2}^{N} \|y_i(t)\|^2.
\end{aligned} \tag{15}
$$

Thus, obtaining an analytical expression for $y_i(t)$ from the recursive formula (14) is key to establishing the consensus property of FedCET. To address this, we introduce Lemma 1.

Lemma 1. *The recursive sequence* $\{a(t)\}_{t=0}^{\infty}$ *satisfying*

$$
a(t+1) = \rho(t)(2a(t) - a(t-1) + b(t) - b(t-1)), \tag{16}
$$

with initial values $a(0)$ *and* $a(1)$*, where*

$$
\rho(t) = \begin{cases} \rho, & t = k\tau, \\ 1, & t \neq k\tau, \end{cases} \tag{17}
$$

for any $\frac{\tau-1}{\tau+3} < \rho < 1$, $k \in \mathbb{Z}$ *and* $1 \leq p \leq \tau$*, we have*

$$
a(k\tau + p) = pa(k\tau + 1) - (p-1)a(k\tau) + \sum_{h=1}^{p-1}\sum_{j=1}^{h}\{b(k\tau + j) - b(k\tau + j - 1)\}, \tag{18}
$$

$$
\begin{aligned}
a(k\tau + 1) =&(\sqrt{\rho})^k (F_{11}(k)a(1) + F_{12}(k)a(0)) + \sum_{s=0}^{k-1}(\sqrt{\rho})^{k-1-s}\{F_{11}(k-1-s)G_{11}(s) \\
&+ F_{12}(k-1-s)G_{21}(s)\},
\end{aligned} \tag{19}
$$

$$
\begin{aligned}
a(k\tau) =&(\sqrt{\rho})^k (F_{21}(k)a(1) + F_{22}(k)a(0)) + \sum_{s=0}^{k-1}(\sqrt{\rho})^{k-1-s}\{F_{21}(k-1-s)G_{11}(s) \\
&+ F_{22}(k-1-s)G_{21}(s)\},
\end{aligned} \tag{20}
$$

where $\cos(\theta) = \frac{\rho(\tau+1)+1-\tau}{2\sqrt{\rho}}$, $\sin(\theta) = \frac{\sqrt{4\rho - [\rho(\tau+1)+(1-\tau)]^2}}{2\sqrt{\rho}}$,

$$
\begin{aligned}
F_{11}(s) &= \frac{\sin[(s+1)\theta]}{\sin[\theta]} + \frac{(\tau-1)\sin[s\theta]}{\sqrt{\rho}\sin[\theta]}, \\
F_{12}(s) &= -\frac{\tau\sqrt{\rho}\sin[s\theta]}{\sin[\theta]}, \quad F_{21}(s) = \frac{\tau\sin[s\theta]}{\sqrt{\rho}\sin[\theta]}, \\
F_{22}(s) &= -\frac{\sin[(s-1)\theta]}{\sin[\theta]} - \frac{(\tau-1)\sin[s\theta]}{\sqrt{\rho}\sin[\theta]}, \\
G_{11}(s) &= \rho\sum_{j=1}^{\tau} j[b(s\tau + \tau + 1 - j) - b(s\tau + \tau - j)], \\
G_{21}(s) &= \sum_{j=1}^{\tau}(j-1)[b(s\tau + \tau + 1 - j) - b(s\tau + \tau - j)].
\end{aligned}
$$

Proof. The recursive fomula (16) can be equivalently expressed as the following matrix form:

$$
x(t+1) = A(t)x(t) + B(t)u(t), \tag{21}
$$

where

$$x(t) = \begin{bmatrix} a(t) \\ a(t-1) \end{bmatrix}, \quad A(t) = \begin{bmatrix} 2\rho(t) & -\rho(t) \\ 1 & 0 \end{bmatrix},$$

$$u(t) = \begin{bmatrix} b(t) \\ b(t-1) \end{bmatrix}, \quad B(t) = \begin{bmatrix} \rho(t) & -\rho(t) \\ 0 & 0 \end{bmatrix}.$$

From the lifting methods introduced in [1], the periodic system (21) can be equivalently expressed as the following time-invariant system

$$x((k+1)\tau+1) = F_1 x(k\tau+1) + G_1 u_1(k), \tag{22}$$

where

$$G_1 = [C(1), \cdots, C(j), \cdots, C(\tau)] \in \mathbb{R}^{2\times 2\tau},$$

$$u_1(k) = \left[u^{\mathbf{T}}(k\tau+1), u^{\mathbf{T}}(k\tau+2), \cdots, u^{\mathbf{T}}(k\tau+\tau)\right]^{\mathbf{T}} \in \mathbb{R}^{2\tau}$$

$$F_1 = \prod_{i=0}^{\tau-1} A(\tau-i) = \begin{bmatrix} \rho(\tau+1) & -\rho\tau \\ \tau & -\tau+1 \end{bmatrix} \in \mathbb{R}^{2\times 2},$$

$$C(j) = \begin{bmatrix} \rho(\tau-j+1) & -\rho(\tau-j+1) \\ \tau-j & j-\tau \end{bmatrix} \in \mathbb{R}^{2\times 2}.$$

From (22), we have

$$x(k\tau+1) = F_1^k x(1) + \sum_{s=0}^{k-1} F_1^{k-1-s} G_1 u_1(s). \tag{23}$$

Then, we establish the properties of F_1^s and $G_1 u_1(s)$ for $s = 1, 2, \cdots, k$ in Lemma 2 (see proof in Appendix 2.1).

Lemma 2. *For $s = 1, 2, \cdots, k$, the matrices F_1^s and $G_1 u_1(s)$ in (23) satisfy*

$$F_1^s = \rho^{\frac{s}{2}} \begin{bmatrix} F_{11}(s) & F_{12}(s) \\ F_{21}(s) & F_{22}(s) \end{bmatrix}, \quad G_1 u_1(s) = \begin{bmatrix} G_{11}(s) \\ G_{21}(s) \end{bmatrix},$$

where $F_{11}(s)$, $F_{12}(s)$, $F_{21}(s)$, $F_{22}(s)$, $G_{11}(s)$, and $G_{21}(s)$, are defined in Lemma 1.

From (23) and Lemma 2, we have

$$\begin{aligned} \begin{bmatrix} a(k\tau+1) \\ a(k\tau) \end{bmatrix} =& (\sqrt{\rho})^k \begin{bmatrix} F_{11}(k) & F_{12}(k) \\ F_{21}(k) & F_{22}(k) \end{bmatrix} \begin{bmatrix} a(1) \\ a(0) \end{bmatrix} \\ &+ \sum_{s=0}^{k-1} (\sqrt{\rho})^{k-1-s} \begin{bmatrix} F_{11}(k-1-s)G_{11}(s) + F_{12}(k-1-s)G_{21}(s) \\ F_{21}(k-1-s)G_{11}(s) + F_{22}(k-1-s)G_{21}(s) \end{bmatrix}. \end{aligned}$$

Thus, (19) and (20) in Lemma 1 can be obtained from the above matrix equation. From (16) and (17), we have

$$\begin{aligned} & a(k\tau+p) - a(k\tau+p-1) \\ =& a(k\tau+p-1) - a(k\tau+p-2) + b(k\tau+p-1) - b(k\tau+p-2) \\ =& a(k\tau+1) - a(k\tau) + \sum_{j=1}^{p-1} \{b(k\tau+j) - b(k\tau+j-1)\}. \end{aligned}$$

for $1 \leq p \leq \tau$. Thus, we have

$$a(k\tau+p) - a(k\tau+1) = (p-1)\{a(k\tau+1) - a(k\tau)\} + \sum_{h=1}^{p-1} \sum_{j=1}^{h} \{b(k\tau+j) - b(k\tau+j-1)\}.$$

The proof of Lemma 1 is completed. □

From (10) and Lemma 1, we can derive the analytical expression for $y_i(t)$ from the recursive formula (14). Thus, for $2 \le i \le N$, we have

$$
\begin{aligned}
&\|y_i(k\tau+1)\| \\
\le&\alpha(\rho\tau+\tau-1)A_1\sum_{s=0}^{k-1}(\sqrt{\rho})^{k-1-s}\{\sum_{j=1}^{\tau}\|h_i(\tau s+1+\tau-j)-h_i(\tau s+\tau-j)\|\}+A_1A_2(\sqrt{\rho})^k,
\end{aligned}
\tag{24}
$$

$$
\begin{aligned}
&\|y_i(k\tau+1)-y_i(k\tau)\| \\
\le&\alpha(\rho\tau+\tau-1)A_3\sum_{s=0}^{k-1}(\sqrt{\rho})^{k-1-s}\{\sum_{j=1}^{\tau}\|h_i(s\tau+\tau+1-j)-h_i(s\tau+\tau-j)\|\}+A_3A_2(\sqrt{\rho})^k,
\end{aligned}
\tag{25}
$$

where

$$
\begin{aligned}
\rho =&1-\xi,\\
A_1 =&\frac{2\tau}{\sqrt{4\rho-[\rho(\tau+1)+1-\tau]^2}},\\
A_2 =&\max_{2\le i\le N}\{\|y_i(1)\|+\|y_i(0)\|\},\\
A_3 =&\max\{\frac{2\sqrt{\rho}+2}{\sqrt{4\rho-[\rho(\tau+1)+(1-\tau)]^2}},\frac{2\sqrt{\rho}}{\sqrt{4\rho-[\rho(\tau+1)+(1-\tau)]^2}}(1+|\tau\sqrt{\rho}-\frac{\tau-1}{\sqrt{\rho}}|)\}.
\end{aligned}
$$

From Lemma 1, (24), and (25), we have

$$
\begin{aligned}
&\|y_i(k\tau+p)\| \\
\le&\|y_i(k\tau+1)\|+(p-1)\|y_i(k\tau+1)-y_i(k\tau)\|+\alpha\sum_{h=1}^{p-1}\sum_{j=1}^{h}\|h_i(k\tau+j)-h_i(k\tau+j-1)\| \\
\le&pA_4A_2(\sqrt{\rho})^k+\alpha\sum_{h=1}^{p-1}\sum_{j=1}^{h}\|h_i(k\tau+j)-h_i(k\tau+j-1)\| \\
&+A_4p\alpha(\rho\tau+\tau-1)\sum_{s=1}^{k}(\sqrt{\rho})^{k-s}\sum_{j=1}^{\tau}\|h_i(s\tau-j)-h_i(s\tau+1-j)\|,
\end{aligned}
\tag{26}
$$

where $A_4=\max\{A_1,A_3\}$ and $1\le p\le\tau$. In FedCET, we consider the $K+1$ communication rounds and the total iteration times satisfy $T=\tau+K\tau$. Thus, we have

$$
\sum_{t=1}^{T}\|y_i(t)\|^2=\sum_{k=0}^{K}\sum_{p=1}^{\tau}\|y_i(k\tau+p)\|^2. \tag{27}
$$

From (26) and (27), we can obtain the upper bound of $\sum_{t=1}^{T}\|y_i(t)\|^2$, which is presented in Lemma 3.

Lemma 3. *There exist* $B_2=3\tau^4+\frac{\tau^2(\tau+1)(2\tau+1)(\rho\tau+\tau-1)^2}{2(1-\sqrt{\rho})^2}A_4^2$ *and* $B_1=\frac{\tau(\tau+1)(2\tau+1)}{2(1-\rho)}A_4^2A_2^2$ *such that*

$$
\sum_{t=1}^{T}\|y_i(t)\|^2\le B_1+\alpha^2B_2\sum_{t=1}^{T-1}\|h_i(t)-h_i(t-1)\|^2.
$$

In addition, we also have

$$
\sum_{i=2}^{N}\|h_i(t-1)-h_i(t)\|^2\le L^2\sum_{i=1}^{N}\|y_i(t)-y_i(t-1)\|^2.
$$

For clarity and readability, we provide the detailed proof of Lemma 3 in the Appendix 2.2. From Lemma 3, we have

$$\sum_{i=2}^{N}\sum_{t=1}^{T}\|y_i(t)\|^2 \leq NB_1 + \alpha^2 L^2 B_2 \sum_{t=1}^{T-1}\sum_{i=1}^{N}\|y_i(t) - y_i(t-1)\|^2. \tag{28}$$

Then, we need to analyze $\|y_1(t) - y_1(t-1)\|^2$. We have $y_1(t) = X(t)Pe_1 = X(t)v_1 = \frac{1}{\sqrt{N}}X(t)\mathbf{1}_N = \sqrt{N}\overline{X}(t)$. Thus, from (6), we have

$$\|y_1(t+1) - y_1(t)\|^2 \leq N\alpha^2\|\overline{\nabla f}(X(t))\|^2. \tag{29}$$

From (28) and (29), we have

$$\begin{aligned}
&\sum_{i=2}^{N}\sum_{t=1}^{T}\|y_i(t)\|^2 \\
\leq& NB_1 + \alpha^2 L^2 B_2 \sum_{t=1}^{T-1}\sum_{i=1}^{N}\|y_i(t) - y_i(t-1)\|^2 \\
\leq& NB_1 + 4\alpha^2 L^2 B_2 \sum_{t=1}^{T-1}\sum_{i=2}^{N}\|y_i(t)\|^2 + 2\alpha^2 L^2 B_2 N A_2^2 + \alpha^4 L^2 N B_2 \sum_{t=1}^{T-1}\|\overline{\nabla f}(X(t-1))\|^2 \\
\leq& C_1 + \alpha^2 L^2 B_2 \sum_{t=1}^{T-1}\left\{4\sum_{i=2}^{N}\|y_i(t)\|^2 + \alpha^2 N\|\overline{\nabla f}(X(t))\|^2\right\}.
\end{aligned}$$

where $C_1 = \alpha^2 L^2 N B_2(\alpha^2\|\overline{\nabla f}(X(0))\|^2 + 2A_2^2) + NB_1$. Combining with (15), we have

$$(1 - 4\alpha^2 L^2 B_2)\sum_{i=1}^{N}\sum_{t=1}^{T}\|\overline{X}(t) - x_i(t)\|^2 \leq C_1 + \alpha^4 L^2 N B_2 \sum_{t=1}^{T-1}\|\overline{\nabla f}(X(t))\|^2, \tag{30}$$

Thus, we obtain the property of the consensus error $\sum_{i=1}^{N}\sum_{t=1}^{T}\|\overline{X}(t) - x_i(t)\|^2$ in (30), which will be utilized to prove the convergence rate of FedCET.

If the stepsize α satisfies $B_3 = 1 - 4\alpha^2 L^2 B_2 > 0$, from (8) and (30), we have

$$\begin{aligned}
&\sum_{t=1}^{T}\{\|\nabla f(\overline{X}(t))\|^2 + (1 - L\alpha - \frac{\alpha^4 L^4 B_2}{B_3})\|\overline{\nabla f}(X(t))\|^2\} \\
\leq& \frac{2}{\alpha}\{f(\overline{X}(1)) - f(x^*)\} + \frac{C_1 L^2}{B_3 N}.
\end{aligned} \tag{31}$$

As for the stepsize $0 < \alpha \leq \frac{1}{\sqrt{5B_2}L}$, we have $1 - L\alpha - \frac{\alpha^4 L^4 B_2}{B_3} > 0$. Thus, from (31), we have

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(\overline{X}(t))\|^2 \leq \frac{2}{\tau\alpha K}\{f(\overline{X}(1)) - f(x^*)\} + \frac{C_1 L^2}{B_3 N\tau K}. \tag{32}$$

Then, we need to analyze the term $\frac{C_1 L^2}{B_3 N(\tau+1)}$ in (32).

$$\begin{aligned}
\frac{C_1 L^2}{B_3 N\tau} \leq& \frac{\alpha^2 L^4 B_2(\alpha^2\|\overline{\nabla f}(X(0))\|^2 + 2A_2^2) + B_1 L^2}{B_3\tau} \\
\leq& \frac{\alpha^2 L^4 B_2(\alpha^2\|\overline{\nabla f}(X(0))\|^2 + 2A_2^2)}{B_3\tau} + \frac{B_1 L^2}{B_3\tau}.
\end{aligned} \tag{33}$$

From the stepsize $0 < \alpha \leq \frac{1}{\sqrt{5B_2}L}$, we have $B_3 = 1 - 4\alpha^2 L^2 B_2 \geq \alpha^2 L^2 B_2$. Thus, we have

$$\frac{B_2}{B_3} \leq \frac{1}{\alpha^2 L^2}, \tag{34}$$

In addition, we also have $A_2^2 \le 2\|X(0)\|_F^2 + 2\|X(1)\|_F^2$, $B_2 = 3\tau^4 + \frac{\tau^2(\tau+1)(2\tau+1)(\rho\tau+\tau-1)^2 A_4^2}{2(1-\sqrt{\rho})^2}$ and $B_1 = \frac{\tau(\tau+1)(2\tau+1)A_4^2 A_2^2}{2(1-\rho)}$. We have

$$B_2 \ge \frac{\rho^2\tau^4(\tau+1)(2\tau+1)A_4^2}{2(1-\sqrt{\rho})^2}.$$

Thus, from (34), we have

$$\begin{aligned}\frac{B_1}{B_3} \le & \frac{1}{\alpha^2 L^2}\frac{B_1}{B_2} \\ \le & \frac{1}{\alpha^2 L^2}\frac{2(1-\sqrt{\rho})^2\tau(\tau+1)(2\tau+1)A_4^2 A_2^2}{2(1-\rho)\rho^2\tau^4(\tau+1)(2\tau+1)A_4^2} \\ \le & \frac{2\|X(0)\|_F^2 + 2\|X(1)\|_F^2}{\alpha^2 L^2 \rho^2 \tau^3}.\end{aligned} \tag{35}$$

From (32), (33), (34), (35), and the stepsize $0 < \alpha \le \frac{1}{\sqrt{5B_2}L}$, we have

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(\overline{X}(t))\|^2 \\ \le & \frac{2}{\tau\alpha K}\{f(\overline{X}(1)) - f(x^*)\} + \frac{\|\overline{\nabla f}(X(0))\|^2}{\tau^3 K} \\ & + \frac{4\|X(0)\|_F^2 + 4\|X(1)\|_F^2}{\tau^3\alpha^2 K} + \frac{2\|X(0)\|_F^2 + 2\|X(1)\|_F^2}{\alpha^2\rho^2\tau^4 K},\end{aligned}$$

where $T = \tau K + \tau$. The proof of Theorem 1 is completed.

2.1 Proof of Lemma 2

The eigenvalues λ_1 and λ_2 of the matrix F_1 satisfy

$$\lambda^2 - [\rho(\tau+1) + (1-\tau)]\lambda + \rho = 0. \tag{36}$$

As for the quadratic equation (36), we have

$$[\rho(\tau+1) + (1-\tau)]^2 - 4\rho < 0,$$

since the parameter ρ satisfies $\rho > \frac{\tau-1}{\tau+3}$. Thus, eigenvalues λ_1 and λ_2 are

$$\begin{aligned}\lambda_1 &= \frac{\rho(\tau+1) + (1-\tau)}{2} + \frac{\sqrt{4\rho - [\rho(\tau+1) + (1-\tau)]^2}}{2}i \\ \lambda_2 &= \frac{\rho(\tau+1) + (1-\tau)}{2} - \frac{\sqrt{4\rho - [\rho(\tau+1) + (1-\tau)]^2}}{2}i.\end{aligned}$$

Thus, λ_1 and λ_2 can be expressed as

$$\lambda_1 = \sqrt{\rho}\{\cos(\theta) + \sin(\theta)i\},\ \lambda_2 = \sqrt{\rho}\{\cos(\theta) + \sin(-\theta)i\},$$

where

$$\begin{cases}\sin(\theta) = & \dfrac{\sqrt{4\rho - [\rho(\tau+1) + (1-\tau)]^2}}{2\sqrt{\rho}}, \\ \cos(\theta) = & \dfrac{\rho(\tau+1) + 1 - \tau}{2\sqrt{\rho}}.\end{cases} \tag{37}$$

Moreover, from Euler's formula, we have eigenvalues $\lambda_1 = \sqrt{\rho}e^{i\theta}$ and $\lambda_2 = \sqrt{\rho}e^{-i\theta}$ with the corresponding eigenvector

$$v_1 = \begin{bmatrix} \frac{\sqrt{\rho}e^{i\theta}+\tau-1}{\tau} \\ 1 \end{bmatrix}, \quad v_2 = \begin{bmatrix} \frac{\sqrt{\rho}e^{-i\theta}+\tau-1}{\tau} \\ 1 \end{bmatrix}.$$

Thus, we can obtain

$$F_1^s = [v_1, v_2] \begin{bmatrix} \lambda_1^s & 0 \\ 0 & \lambda_2^s \end{bmatrix} [v_1, v_2]^{-1}$$

for any $s = 0, 1, \cdots, k$, where

$$[v_1, v_2]^{-1} = \frac{-i\tau}{2\sqrt{\rho}\sin(\theta)} \begin{bmatrix} 1 & -\frac{\sqrt{\rho}e^{-i\theta}+\tau-1}{\tau} \\ -1 & \frac{\sqrt{\rho}e^{i\theta}+\tau-1}{\tau} \end{bmatrix}.$$

Then, we know that for any $s \geq 0$, we have

$$\begin{aligned} F_1^s =& [v_1, v_2] \begin{bmatrix} \lambda_1^s & 0 \\ 0 & \lambda_2^s \end{bmatrix} [v_1, v_2]^{-1} \\ =& \frac{-i\tau(\sqrt{\rho})^s}{2\sqrt{\rho}\sin(\theta)} \begin{bmatrix} \frac{\sqrt{\rho}e^{i\theta}+\tau-1}{\tau} & \frac{\sqrt{\rho}e^{-i\theta}+\tau-1}{\tau} \\ 1 & 1 \end{bmatrix} \begin{bmatrix} e^{s\theta i} & 0 \\ 0 & e^{-s\theta i} \end{bmatrix} \begin{bmatrix} 1 & -\frac{\sqrt{\rho}e^{-i\theta}+\tau-1}{\tau} \\ -1 & \frac{\sqrt{\rho}e^{i\theta}+\tau-1}{\tau} \end{bmatrix} \\ =& \frac{-i\tau(\sqrt{\rho})^s}{2\sqrt{\rho}\sin(\theta)} \begin{bmatrix} \frac{\sqrt{\rho}e^{i\theta}+\tau-1}{\tau} & \frac{\sqrt{\rho}e^{-i\theta}+\tau-1}{\tau} \\ 1 & 1 \end{bmatrix} \begin{bmatrix} e^{s\theta i} & -\frac{\sqrt{\rho}e^{-i\theta}+\tau-1}{\tau}e^{s\theta i} \\ -e^{-s\theta i} & \frac{\sqrt{\rho}e^{i\theta}+\tau-1}{\tau}e^{-s\theta i} \end{bmatrix} \\ =& (\sqrt{\rho})^s \begin{bmatrix} F_{11}(s) & F_{12}(s) \\ F_{21}(s) & F_{22}(s) \end{bmatrix}. \end{aligned}$$

As for the $(1,1)^{th}$ element of F_1^s, we have

$$\begin{aligned} & F_{11}(s) \\ =& \frac{-i\tau}{2\sqrt{\rho}\sin(\theta)} \Big\{ \frac{\sqrt{\rho}e^{i\theta}+\tau-1}{\tau} e^{s\theta i} - \frac{\sqrt{\rho}e^{-i\theta}+\tau-1}{\tau} e^{-s\theta i} \Big\} \\ =& \frac{-i\tau}{2\sqrt{\rho}\sin(\theta)} \Big\{ \frac{\sqrt{\rho}}{\tau}(e^{i(s+1)\theta} - e^{-i(s+1)\theta}) + \frac{\tau-1}{\tau}(e^{is\theta} - e^{-is\theta}) \Big\} \\ =& \frac{-i\tau}{2\sqrt{\rho}\sin(\theta)} \Big\{ \frac{2\sqrt{\rho}}{\tau} i\sin[(s+1)\theta] + \frac{2(\tau-1)}{\tau} i\sin[s\theta] \Big\} \\ =& \frac{\sin[(s+1)\theta]}{\sin[\theta]} + \frac{(\tau-1)\sin[s\theta]}{\sqrt{\rho}\sin[\theta]} \end{aligned}$$

As for the $(2,1)^{th}$ element of F_1^s, we have

$$F_{21}(s) = \frac{-i\tau}{2\sqrt{\rho}\sin(\theta)} \{e^{s\theta i} - e^{-s\theta i}\} = \frac{\tau\sin[s\theta]}{\sqrt{\rho}\sin[\theta]}.$$

As for the $(2,2)^{th}$ element of F_1^s, we have

$$\begin{aligned} & F_{22}(s) \\ =& \frac{-i\tau}{2\sqrt{\rho}\sin(\theta)} \Big\{ \frac{\sqrt{\rho}e^{i\theta}+\tau-1}{\tau} e^{-s\theta i} - \frac{\sqrt{\rho}e^{-i\theta}+\tau-1}{\tau} e^{s\theta i} \Big\} \\ =& \frac{-i\tau}{2\sqrt{\rho}\sin(\theta)} \Big\{ \frac{\sqrt{\rho}}{\tau}(e^{-i(s-1)\theta} - e^{i(s-1)\theta}) + \frac{\tau-1}{\tau}(e^{-is\theta} - e^{is\theta}) \Big\} \\ =& \frac{-i\tau}{2\sqrt{\rho}\sin(\theta)} \Big\{ -\frac{2\sqrt{\rho}}{\tau} i\sin[(s-1)\theta] - \frac{2(\tau-1)}{\tau} i\sin[s\theta] \Big\} \\ =& -\frac{\sin[(s-1)\theta]}{\sin[\theta]} - \frac{(\tau-1)\sin[s\theta]}{\sqrt{\rho}\sin[\theta]} \end{aligned}$$

As for the $(1,2)^{th}$ element of F_1^s, from (37), we have

$$
\begin{aligned}
&F_{12}(s)\\
=&\frac{-i\tau}{2\sqrt{\rho}\sin(\theta)}\Big\{\frac{\sqrt{\rho}e^{-i\theta}+\tau-1}{\tau}\frac{\sqrt{\rho}e^{i\theta}+\tau-1}{\tau}e^{-s\theta i}-\frac{\sqrt{\rho}e^{i\theta}+\tau-1}{\tau}\frac{\sqrt{\rho}e^{-i\theta}+\tau-1}{\tau}e^{s\theta i}\Big\}\\
=&\frac{-i\tau}{2\sqrt{\rho}\sin(\theta)}\Big\{-\frac{\sqrt{\rho}e^{-i\theta}+\tau-1}{\tau}\frac{\sqrt{\rho}e^{i\theta}+\tau-1}{\tau}2i\sin[s\theta]\Big\}\\
=&-\frac{\sin(s\theta)}{\tau\sqrt{\rho}\sin(\theta)}\Big((\tau-1+\sqrt{\rho}\cos[\theta])^2+\rho\sin^2[\theta]\Big)\\
=&-\frac{\sin(s\theta)}{\tau\sqrt{\rho}\sin(\theta)}\Big((\tau-1)^2+2\sqrt{\rho}\cos[\theta](\tau-1)+\rho\Big)\\
=&-\frac{\sqrt{\rho}\tau\sin(s\theta)}{\sin(\theta)}.
\end{aligned}
$$

Based on the definition of G_1 and $u_1(s)$, we can obtain

$$
\begin{aligned}
G_{11}(s)&=\rho\sum_{j=1}^{\tau}j[b(s\tau+\tau+1-j)-b(s\tau+\tau-j)],\\
G_{21}(s)&=\sum_{j=1}^{\tau}(j-1)[b(s\tau+\tau+1-j)-b(s\tau+\tau-j)].
\end{aligned}
$$

The proof of Lemma 2 is completed.

2.2 Proof of Lemma 3

From (26), we have

$$
\begin{aligned}
&\|y_i(k\tau+p)\|^2\\
\le&3p^2A_4^2A_2^2\rho^k+3\alpha^2\Big(\sum_{h=1}^{p-1}\sum_{j=1}^{h}\|h_i(k\tau+j)-h_i(k\tau+j-1)\|\Big)^2\\
&+3A_4^2p^2\alpha^2(\rho\tau+\tau-1)^2\Big(\sum_{s=1}^{k}(\sqrt{\rho})^{k-s}\sum_{j=1}^{\tau}\|h_i(s\tau+1-j)-h_i(s\tau-j)\|\Big)^2. \qquad (38)
\end{aligned}
$$

Then, we need the following lemma from [2] to analyze the three terms of $\sum_{t=1}^{T}\|y_i(t)\|^2=\sum_{k=0}^{K}\sum_{p=1}^{\tau}\|y_i(k\tau+p)\|^2$ in (38).

Lemma 4 ([2]). *For two non-negative sequences $\{a(t)\}_{t=1}^{\infty}$ and $\{b(t)\}_{t=1}^{\infty}$ satisfying $a(t)=\sum_{s=1}^{t}\rho^{t-s}b(s)$ with $0\le\rho<1$, we have*

$$
\sum_{t=s}^{k}a(t)\le\sum_{t=s}^{k}\frac{b(s)}{1-\rho},\qquad \sum_{t=s}^{k}a^2(t)\le\sum_{t=s}^{k}\frac{b^2(s)}{(1-\rho)^2}.
$$

It is worth noting that $\sum_{p=1}^{\tau}p^2=\frac{\tau(\tau+1)(2\tau+1)}{6}$. Thus, as for the first term of (38), we have

$$
\sum_{k=0}^{K}\sum_{p=1}^{\tau}3p^2A_4^2A_2^2\rho^k\le\frac{\tau(\tau+1)(2\tau+1)A_4^2A_2^2}{2(1-\rho)}. \qquad (39)
$$

As for the second term of (38), we have

$$
\begin{aligned}
&\sum_{k=0}^{K}\sum_{p=1}^{\tau}\alpha^2(\sum_{h=1}^{p-1}\sum_{j=1}^{h}\|h_i(k\tau+j)-h_i(k\tau+j-1)\|)^2\\
\leq&\alpha^2\sum_{k=0}^{K}\sum_{p=1}^{\tau}(\sum_{h=1}^{\tau-1}\sum_{j=1}^{h}\|h_i(k\tau+j)-h_i(k\tau+j-1)\|)^2\\
\leq&\alpha^2\tau^2\sum_{k=0}^{K}\sum_{p=1}^{\tau}(\sum_{j=1}^{\tau-1}\|h_i(k\tau+j)-h_i(k\tau+j-1)\|)^2\\
\leq&\alpha^2\tau^3\sum_{k=0}^{K}(\sum_{j=1}^{\tau-1}\|h_i(k\tau+j)-h_i(k\tau+j-1)\|)^2\\
\leq&\alpha^2\tau^4\sum_{k=0}^{K}\sum_{j=1}^{\tau-1}\|h_i(k\tau+j)-h_i(k\tau+j-1)\|^2.
\end{aligned}\tag{40}
$$

As for the third term of (38), from Lemma 4, we have

$$
\begin{aligned}
&\sum_{k=1}^{K}(\sum_{s=1}^{k}(\sqrt{\rho})^{k-s}\sum_{j=1}^{\tau}\|h_i(\tau s+1-j)-h_i(\tau s-j)\|)^2\\
\leq&\frac{\sum_{k=1}^{K}(\sum_{j=1}^{\tau}\|h_i(k\tau+1-j)-h_i(k\tau-j)\|)^2}{(1-\sqrt{\rho})^2}\\
\leq&\frac{\tau\sum_{k=1}^{K}\sum_{j=1}^{\tau}\|h_i(k\tau+1-j)-h_i(k\tau-j)\|^2}{(1-\sqrt{\rho})^2}.
\end{aligned}\tag{41}
$$

Thus, from (38), (39), (41), and (40), we have

$$
\sum_{t=1}^{T}\|y_i(t)\|^2\leq B_1+\alpha^2B_2\sum_{t=1}^{T-1}\|h_i(t)-h_i(t-1)\|^2.
$$

As for the term $\|h_i(t)-h_i(t-1)\|^2$ in above equation, from (12) and Assumption 1‘, we have

$$
\begin{aligned}
&\sum_{i=2}^{N}\|h_i(t-1)-h_i(t)\|^2\\
\leq&\sum_{i=1}^{N}\|\overline{\nabla f}(X(t))Pe(i)-\overline{\nabla f}(X(t-1))Pe(i)\|^2\\
=&\|\overline{\nabla f}(X(t))-\overline{\nabla f}(X(t-1))\|_F^2\\
=&\sum_{i=1}^{N}\|\nabla f_i(x_i(t))-\nabla f_i(x_i(t-1))\|^2\\
\leq&L^2\sum_{i=1}^{N}\|x_i(t)-x_i(t-1)\|^2\\
=&L^2\sum_{i=1}^{N}\|Y(t)P^{\mathbf{T}}e(i)-Y(t-1)P^{\mathbf{T}}e(i)\|^2\\
=&L^2\sum_{i=1}^{N}\|y_i(t)-y_i(t-1)\|^2.
\end{aligned}
$$

The proof of Lemma 3 is completed.

3 Proof of Theorem 2

Based on our assumptions, the stochastic gradient $\nabla f_i(x,\xi_i)$ is an unbiased estimate of the accurate gradient $\nabla f_i(x)$, with its variance bounded by σ^2. Thus, we have

$$\mathbb{E}_{\xi_i\sim D_i}[\nabla f_i(x,\xi_i)]=\nabla f_i(x), \quad \mathbb{E}_{\xi_i\sim D_i}[\|\nabla f_i(x,\xi_i)-\nabla f_i(x)\|^2]\le\sigma^2, \tag{42}$$

for any $x\in\mathbb{R}^n$ and $i\in\mathcal{S}$, where $\xi_i(t)\sim D_i$ are samples drawn from the local data distribution at each iteration. We define

$$\begin{aligned}
G(t)=&[\nabla f_1(x_1(t),\xi_1(t)),\cdots,\nabla f_N(x_N(t),\xi_N(t))],\\
\overline{G}(t)=&\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i(t),\xi_i(t)).
\end{aligned}$$

Similarly to (4), we have

$$\overline{X}(t+1)=\overline{X}(t)-\alpha\overline{G}(t).$$

From Assumption 1 and (42), we have

$$\begin{aligned}
&\mathbb{E}[f(\overline{X}(t+1))]\\
\le&\mathbb{E}[f(\overline{X}(t))]-\mathbb{E}\langle\nabla f(\overline{X}(t)),\alpha\overline{G}(t)\rangle+\frac{L\alpha^2}{2}\mathbb{E}\|\overline{G}(t)\|^2\\
=&\mathbb{E}[f(\overline{X}(t))]-\alpha\mathbb{E}\langle\nabla f(\overline{X}(t)),\overline{\nabla f}(X(t))\rangle+\frac{L\alpha^2}{2}\mathbb{E}\|\overline{\nabla f}(X(t))\|^2\\
&+\frac{L\alpha^2}{2}\mathbb{E}\|\overline{G}(t)-\overline{\nabla f}(X(t))\|^2+L\alpha^2\mathbb{E}[\langle\overline{G}(t)-\overline{\nabla f}(X(t)),\overline{\nabla f}(X(t))\rangle]\\
=&\mathbb{E}[f(\overline{X}(t))]-\alpha\mathbb{E}\langle\nabla f(\overline{X}(t)),\overline{\nabla f}(X(t))\rangle\\
&+\frac{L\alpha^2}{2}\mathbb{E}\|\overline{\nabla f}(X(t))\|^2+\frac{L\alpha^2}{2}\mathbb{E}\|\overline{G}(t)-\overline{\nabla f}(X(t))\|^2\\
\le&\mathbb{E}[f(\overline{X}(t))]-\alpha\mathbb{E}\langle\nabla f(\overline{X}(t)),\overline{\nabla f}(X(t))\rangle+\frac{L\alpha^2}{2}\mathbb{E}\|\overline{\nabla f}(X(t))\|^2\\
&+\frac{L\alpha^2}{2N}\sum_{i=1}^{N}\mathbb{E}\|\nabla f_i(x_i(t),\xi_i(t))-\nabla f_i(x_i(t))\|^2\\
\le&\mathbb{E}[f(\overline{X}(t))]-\alpha\mathbb{E}\langle\nabla f(\overline{X}(t)),\overline{\nabla f}(X(t))\rangle+\frac{L\alpha^2}{2}\mathbb{E}\|\overline{\nabla f}(X(t))\|^2+\frac{L\alpha^2}{2}\sigma^2\\
=&\mathbb{E}[f(\overline{X}(t))]-\frac{\alpha}{2}\mathbb{E}\|\nabla f(\overline{X}(t))\|^2-(\frac{\alpha}{2}-\frac{L\alpha^2}{2})\mathbb{E}\|\overline{\nabla f}(X(t))\|^2\\
&+\frac{\alpha}{2}\mathbb{E}\|\nabla f(\overline{X}(t))-\overline{\nabla f}(X(t))\|^2+\frac{L\alpha^2}{2}\sigma^2.
\end{aligned}$$

Then, we have

$$\begin{aligned}
&\mathbb{E}\|\nabla f(\overline{X}(t))-\overline{\nabla f}(X(t))\|^2\\
=&\mathbb{E}\|\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(\overline{X}(t))-\frac{1}{N}\sum_{i=1}^{N}\nabla f_i(x_i(t))\|^2\\
\le&\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\|\nabla f_i(\overline{X}(t))-\nabla f_i(x_i(t))\|^2\\
\le&\frac{L^2}{N}\sum_{i=1}^{N}\mathbb{E}\|\overline{X}(t)-x_i(t)\|^2.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&\frac{\alpha}{2}\mathbb{E}\|\nabla f(\overline{X}(t))\|^2+(\frac{\alpha}{2}-\frac{L\alpha^2}{2})\mathbb{E}\|\overline{\nabla f}(X(t))\|^2\\
\le&\mathbb{E}f(\overline{X}(t))-\mathbb{E}f(\overline{X}(t+1))+\frac{\alpha L^2}{2N}\sum_{i=1}^{N}\mathbb{E}\|\overline{X}(t)-x_i(t)\|^2+\frac{L\alpha^2}{2}\sigma^2.
\end{aligned}\tag{43}$$

Then, we need the following Lemma 5, which is proved in Appendix 3.1.

Lemma 5. *We define that* $B_4 = 1 - 12\alpha^2 L^2 B_2$ *and we have*

$$\begin{aligned}&\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbb{E}[\|\overline{X}(t) - x_i(t)\|^2] \\ \le&\frac{C_2}{B_4 NT} + \frac{6\alpha^4 L^2 B_2}{B_4 T}\sum_{t=1}^{T-1}\mathbb{E}[\|\overline{\nabla f}(X(t))\|^2] + \frac{12B_2}{B_4}\alpha^2\sigma^2,\end{aligned}$$

where $C_2 = 6\alpha^2 L^2 N B_2(\alpha^2\mathbb{E}[\|\overline{\nabla f}(X(0))\|^2] + A_5^2) + NB_1$ *and* $A_5 = 2\mathbb{E}[\|X(1)\|^2] + 2\mathbb{E}[\|X(0)\|^2]$.

From (43) and Lemma 5, we have

$$\begin{aligned}&\sum_{t=1}^{T}\left\{\mathbb{E}\|\nabla f(\overline{X}(t))\|^2 + (1 - L\alpha)\mathbb{E}\|\overline{\nabla f}(X(t))\|^2\right\} \\ \le&\frac{2}{\alpha}\left\{\mathbb{E}[f(\overline{X}(1))] - f(x^*)\right\} + L\alpha\sigma^2 T + \frac{L^2}{N}\sum_{t=1}^{T}\sum_{i=1}^{N}\mathbb{E}\|\overline{X}(t) - x_i(t)\|^2 \\ \le&\frac{2}{\alpha}\left\{\mathbb{E}[f(\overline{X}(1))] - f(x^*)\right\} + L\alpha\sigma^2 T + \frac{C_2 L^2}{B_4 N} \\ &+ \frac{12B_2 L^2}{B_4}\alpha^2\sigma^2 T + \frac{6\alpha^4 L^4 B_2}{B_4}\sum_{t=1}^{T-1}\mathbb{E}[\|\overline{\nabla f}(X(t))\|^2].\end{aligned}$$

If the stepsize $0 < \alpha \le \frac{1}{\sqrt{13 B_2} L}$, we have $B_4 \ge \alpha^2 L^2 B_2$ and $1 - L\alpha - \frac{6\alpha^4 L^4 B_2}{B_4} > 0$. Thus, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\overline{X}(t))\|^2 \le& \frac{2}{\alpha T}\{\mathbb{E}f(\overline{X}(1)) - f(x^*)\} \\ &+ \frac{C_2 L^2}{B_4 NT} + L\alpha\sigma^2 + \frac{12B_2}{1 - 12\alpha^2 L^2 B_2}\alpha^2 L^2\sigma^2.\end{aligned} \tag{44}$$

Thus, we need to analyze the term $\frac{C_2 L^2}{B_4 NT}$, which satisfies

$$\begin{aligned}&\frac{C_2 L^2}{B_4 N\tau K} \\ \le&\frac{6\alpha^2 L^4 B_2(\alpha^2\mathbb{E}[\|\overline{\nabla f}(X(0))\|^2] + A_5^2) + B_1 L^2}{B_4\tau K} \\ \le&\frac{6\alpha^4 L^4 B_2\mathbb{E}[\|\overline{\nabla f}(X(0))\|^2]}{B_4\tau K} + \frac{6\alpha^2 L^4 B_2 A_5^2}{B_4\tau K} + \frac{B_1 L^2}{B_4\tau K} \\ \le&\frac{6\mathbb{E}[\|\overline{\nabla f}(X(0))\|^2]}{\tau^3 K} + \frac{12\mathbb{E}[\|X(1)\|_F^2] + 12\mathbb{E}[\|X(0)\|_F^2]}{\tau^3\alpha^2 K} + \frac{B_1 L^2}{B_4\tau K}.\end{aligned} \tag{45}$$

Similar to (35), we have

$$\begin{aligned}\frac{B_1}{B_4} \le&\frac{1}{\alpha^2 L^2}\frac{B_1}{B_2} \\ \le&\frac{1}{\alpha^2 L^2}\frac{2(1-\sqrt{\rho})^2\tau(\tau+1)(2\tau+1)A_4^2 A_2^2}{2(1-\rho)\rho^2\tau^4(\tau+1)(2\tau+1)A_4^2} \\ \le&\frac{2\mathbb{E}[\|X(0)\|_F^2] + 2\mathbb{E}[\|X(0)\|^2]}{\alpha^2 L^2\rho^2\tau^3}.\end{aligned} \tag{46}$$

In addition, from $0 < \alpha \le \frac{1}{\sqrt{13 B_2} L}$, we have

$$1 - 12\alpha^2 B_2 L^2 \ge \frac{1}{13}. \tag{47}$$

Thus, from (44), (45), (46), and (47), we have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\overline{X}(t))\|^2\\
\leq&\frac{2\mathbb{E}[f(\overline{X}(1))]-2f(x^*)}{\alpha\tau K}+\frac{2\mathbb{E}[\|X(1)\|_F^2]+2\mathbb{E}[\|X(0)\|_F^2]}{\alpha^2\rho^2\tau^4K}+\frac{6\mathbb{E}[\|\overline{\nabla f}(X(0))\|^2]}{\tau^3K}\\
&+\frac{12\mathbb{E}[\|X(1)\|_F^2]+12\mathbb{E}[\|X(0)\|_F^2]}{\tau^3\alpha^2K}+L\alpha\sigma^2+156B_2\alpha^2L^2\sigma^2.
\end{aligned}$$

The proof of Theorem 2 is completed.

3.1 Proof of Lemma 5

Similar to Lemma 3, we have

$$\sum_{t=1}^{T}\|y_i(t)\|^2\leq B_1+\alpha^2B_2\sum_{t=1}^{T-1}\|h_i(t)-h_i(t-1)\|^2, \tag{48}$$

where

$$\begin{aligned}
B_2=&3\tau^4+\frac{\tau^2(\tau+1)(2\tau+1)(\rho\tau+\tau-1)^2A_4^2}{2(1-\sqrt{\rho})^2},\\
B_1=&\frac{\tau(\tau+1)(2\tau+1)A_4^2A_2^2}{2(1-\rho)}.
\end{aligned}$$

In addition, we have

$$\begin{aligned}
&\sum_{i=2}^{N}\mathbb{E}[\|h_i(t)-h_i(t-1)\|^2]\\
\leq&\sum_{i=1}^{N}\mathbb{E}[\|G(t)Pe(i)-G(t-1)Pe(i)\|^2]\\
=&\mathbb{E}[\|G(t)-G(t-1)\|_F^2]\\
\leq&3\sum_{i=1}^{N}\mathbb{E}[\|\nabla f_i(x_i(t-1),\xi_i(t-1))-\nabla f_i(x_i(t-1))\|^2]\\
&+3\sum_{i=1}^{N}\mathbb{E}[\|\nabla f_i(x_i(t),\xi_i(t))-\nabla f_i(x_i(t))\|^2]+3\sum_{i=1}^{N}\mathbb{E}[\|\nabla f_i(x_i(t))-\nabla f_i(x_i(t-1))\|^2]\\
\leq&6\sigma^2N+3L^2\sum_{i=1}^{N}\mathbb{E}[\|x_i(t)-x_i(t-1)\|^2]\\
=&6\sigma^2N+3L^2\mathbb{E}[\|Y(t)P^{\mathbf{T}}-Y(t-1)P^{\mathbf{T}}\|_F^2]\\
=&6\sigma^2N+3L^2\sum_{i=1}^{N}\mathbb{E}[\|y_i(t)-y_i(t-1)\|^2].
\end{aligned}\tag{49}$$

From (48) and (49), we have

$$\sum_{i=2}^{N}\sum_{t=1}^{T}\mathbb{E}[\|y_i(t)\|^2]\leq NB_1+3\alpha^2L^2B_2\sum_{t=1}^{T-1}\sum_{i=1}^{N}\mathbb{E}[\|y_i(t)-y_i(t-1)\|^2]+6\alpha^2NB_2\sigma^2T. \tag{50}$$

We already have

$$y_1(t)=X(t)Pe_1=X(t)v_1=\sqrt{N}\overline{X}(t).$$

In addition, we have

$$\begin{aligned}
&\|\overline{X}(t+1)-\overline{X}(t)\|^2\\
\leq&2\alpha^2\|\overline{G}(t)-\overline{\nabla f}(X(t))\|^2+2\alpha^2\|\overline{\nabla f}(X(t))\|^2\\
\leq&2\alpha^2\|\frac{1}{N}\sum_{i=1}^N\nabla f_i(x_i(t),\xi_i(t))-\frac{1}{N}\sum_{i=1}^N\nabla f_i(x_i(t))\|^2+2\alpha^2\|\overline{\nabla f}(X(t))\|^2\\
\leq&2\alpha^2\sigma^2+2\alpha^2\mathbb{E}[\|\overline{\nabla f}(X(t))\|^2]. \qquad (51)
\end{aligned}$$

Thus, from (51), we have

$$\mathbb{E}[\|y_1(t+1)-y_1(t)\|^2]\leq 2\alpha^2N\sigma^2+2\alpha^2N\mathbb{E}[\|\overline{\nabla f}(X(t))\|^2]. \qquad (52)$$

From (50) and (52), we have

$$\begin{aligned}
&\sum_{i=2}^N\sum_{t=1}^T\mathbb{E}[\|y_i(t)\|^2]\\
\leq&NB_1+3\alpha^2L^2B_2\sum_{t=1}^{T-1}\sum_{i=2}^N\mathbb{E}[\|y_i(t)-y_i(t-1)\|^2]\\
&+3\alpha^2L^2B_2\sum_{t=1}^{T-1}\{2\alpha^2N\sigma^2+2\alpha^2N\mathbb{E}[\|\overline{\nabla f}(X(t-1))\|^2]\}+6\alpha^2NB_2\sigma^2T\\
\leq&NB_1+6\alpha^2L^2B_2\sum_{t=1}^{T-1}\sum_{i=2}^N\mathbb{E}[\|y_i(t)\|^2+\|y_i(t-1)\|^2]\\
&+6\alpha^4L^2B_2N\sum_{t=1}^{T-1}\mathbb{E}[\|\overline{\nabla f}(X(t-1))\|^2]+12B_2N\alpha^2\sigma^2T.
\end{aligned}$$

Thus, from (15), we have

$$\begin{aligned}
&(1-12\alpha^2L^2B_2)\sum_{i=1}^N\sum_{t=1}^T\mathbb{E}[\|\overline{X}(t)-x_i(t)\|^2]\\
\leq&C_2+6\alpha^4L^2B_2N\sum_{t=1}^{T-1}\mathbb{E}[\|\overline{\nabla f}(X(t))\|^2]+12B_2N\alpha^2\sigma^2T.
\end{aligned}$$

where $C_2=6\alpha^2L^2NB_2(\alpha^2\mathbb{E}[\|\overline{\nabla f}(X(0))\|^2]+A_5^2)+NB_1$ and $A_5=\max_{i=2,3,\cdots,N}\{\mathbb{E}[\|y_i(1)\|]+\mathbb{E}[\|y_i(0)\|]\}$. Moreover, we define $B_4=1-12\alpha^2L^2B_2$ and the proof of Lemma 5 is completed.

References

[1] Sergio Bittanti and Patrizio Colaneri. Invariant representations of discrete-time periodic systems. *Automatica*, 36(12):1777–1793, 2000.

[2] Hanlin Tang, Xiangru Lian, Ming Yan, Ce Zhang, and Ji Liu. D^2: Decentralized training over decentralized data. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 4848–4856, 2018.