

## 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 exist the stepsize  $\alpha$  and the parameter  $0 < c < \frac{2}{(\tau+3)\alpha}$  such that

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

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 exist the stepsize  $\alpha$  and the parameter  $0 < c < \frac{2}{(\tau+3)\alpha}$  such that

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

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:

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

In addition, we define the time-varying mixing matrix

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

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

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

From (2), we have

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

From (1) and (3), we have

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

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

$$x_i(0) = x_i(-1) - \alpha \nabla f_i(x_i(-1)). \quad (5)$$

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

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

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

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

23 Thus, from (7), we have

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

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

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

27 where

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

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

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

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

$$W(t) = P \Lambda(t) P^T, \quad (11)$$

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

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

33 from (2) and (11), we have

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

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

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

for  $i = 1, 2, \dots, N$ . We denote the  $n$ -dimensional unit vector set as  $\{e_i\}_{i=1,2,\dots,n}$ , 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  $\bar{X}(t) - x_i(t)$  as follows:

$$\begin{aligned}
& \sum_{i=1}^N \|\bar{X}(t) - x_i(t)\|^2 \\
&= \sum_{i=1}^N \|X(t)e_i - \frac{1}{N}X(t)\mathbf{1}_N\|^2 \\
&= \|X(t)v_1v_1^T - X(t)PP^T\|_F^2 \\
&= \|X(t)P(\mathbf{I}_N - [e_1, \mathbf{0}_N, \dots, \mathbf{0}_N])P^T\|_F^2 \\
&= \sum_{i=2}^N \|y_i(t)\|^2.
\end{aligned} \tag{15}$$

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

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

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

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

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

$$F_{11}(s) = \frac{\sin[(s+1)\theta]}{\sin[\theta]} + \frac{(\tau-1)\sin[s\theta]}{\sqrt{\rho}\sin[\theta]},$$

$$F_{12}(s) = -\frac{\tau\sqrt{\rho}\sin[s\theta]}{\sin[\theta]}, \quad F_{21}(s) = \frac{\tau\sin[s\theta]}{\sqrt{\rho}\sin[\theta]},$$

$$F_{22}(s) = -\frac{\sin[(s-1)\theta]}{\sin[\theta]} - \frac{(\tau-1)\sin[s\theta]}{\sqrt{\rho}\sin[\theta]},$$

$$G_{11}(s) = \rho \sum_{j=1}^{\tau} j[b(s\tau + \tau + 1 - j) - b(s\tau + \tau - j)],$$

$$G_{21}(s) = \sum_{j=1}^{\tau} (j-1)[b(s\tau + \tau + 1 - j) - b(s\tau + \tau - j)].$$

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

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

45 where

$$46 \quad \begin{aligned} x(t) &= \begin{bmatrix} a(t) \\ a(t-1) \end{bmatrix}, \quad A(t) = \begin{bmatrix} 2\rho(t) & -\rho(t) \\ 1 & 0 \end{bmatrix}, \\ u(t) &= \begin{bmatrix} b(t) \\ b(t-1) \end{bmatrix}, \quad B(t) = \begin{bmatrix} \rho(t) & -\rho(t) \\ 0 & 0 \end{bmatrix}. \end{aligned}$$

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

$$x((k+1)\tau+1) = F_1 x(k\tau+1) + G_1 u_1(k), \quad (22)$$

49 where

$$\begin{aligned} G_1 &= [C(1), \dots, C(j), \dots, C(\tau)] \in \mathbb{R}^{2 \times 2\tau}, \\ u_1(k) &= [u^{\mathbf{T}}(k\tau+1), u^{\mathbf{T}}(k\tau+2), \dots, u^{\mathbf{T}}(k\tau+\tau)]^{\mathbf{T}} \in \mathbb{R}^{2\tau} \\ 50 \quad F_1 &= \prod_{i=0}^{\tau-1} A(\tau-i) = \begin{bmatrix} \rho(\tau+1) & -\rho\tau \\ \tau & -\tau+1 \end{bmatrix} \in \mathbb{R}^{2 \times 2}, \\ 51 \quad C(j) &= \begin{bmatrix} \rho(\tau-j+1) & -\rho(\tau-j+1) \\ \tau-j & j-\tau \end{bmatrix} \in \mathbb{R}^{2 \times 2}. \end{aligned}$$

52 From (22), we have

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

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

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

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

56 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.

57 From (23) and Lemma 2, we have

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

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

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

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

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

61 The proof of Lemma 1 is completed.  $\square$

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

$$\begin{aligned} & \|y_i(k\tau + 1)\| \\ & \leq \alpha(\rho\tau + \tau - 1)A_1 \sum_{s=0}^{k-1} (\sqrt{\rho})^{k-1-s} \left\{ \sum_{j=1}^{\tau} \|h_i(\tau s + 1 + \tau - j) - h_i(\tau s + \tau - j)\| \right\} + A_1 A_2 (\sqrt{\rho})^k, \end{aligned} \quad (24)$$

$$\begin{aligned} & \|y_i(k\tau + 1) - y_i(k\tau)\| \\ & \leq \alpha(\rho\tau + \tau - 1)A_3 \sum_{s=0}^{k-1} (\sqrt{\rho})^{k-1-s} \left\{ \sum_{j=1}^{\tau} \|h_i(s\tau + \tau + 1 - j) - h_i(s\tau + \tau - j)\| \right\} + A_3 A_2 (\sqrt{\rho})^k, \end{aligned} \quad (25)$$

where

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

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

$$\begin{aligned} & \|y_i(k\tau + p)\| \\ & \leq \|y_i(k\tau + 1)\| + (p - 1)\|y_i(k\tau + 1) - y_i(k\tau)\| + \alpha \sum_{h=1}^{p-1} \sum_{j=1}^h \|h_i(k\tau + j) - h_i(k\tau + j - 1)\| \\ & \leq pA_4 A_2 (\sqrt{\rho})^k + \alpha \sum_{h=1}^{p-1} \sum_{j=1}^h \|h_i(k\tau + j) - h_i(k\tau + j - 1)\| \\ & \quad + A_4 p \alpha (\rho\tau + \tau - 1) \sum_{s=1}^k (\sqrt{\rho})^{k-s} \sum_{j=1}^{\tau} \|h_i(s\tau - j) - h_i(s\tau + 1 - j)\|, \end{aligned} \quad (26)$$

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

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

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  $B_1 = \frac{\tau(\tau+1)(2\tau+1)}{2(1-\rho)} A_4^2 A_2^2$  such that*

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

In addition, we also have

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

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

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

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

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

From (28) and (29), we have

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

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

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

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

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

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

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

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

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

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

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

$$\frac{B_2}{B_3} \leq \frac{1}{\alpha^2 L^2}, \quad (34)$$

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

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

87 Thus, from (34), we have

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

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

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

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

## 90 2.1 Proof of Lemma 2

91 The eigenvalues  $\lambda_1$  and  $\lambda_2$  of the matrix  $F_1$  satisfy

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

92 As for the quadratic equation (36), we have

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

93 since the parameter  $\rho$  satisfies  $\rho > \frac{\tau-1}{\tau+3}$ . Thus, eigenvalues  $\lambda_1$  and  $\lambda_2$  are

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

94 Thus,  $\lambda_1$  and  $\lambda_2$  can be expressed as

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

95 where

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

96 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  
 97 corresponding eigenvector

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

98 Thus, we can obtain

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

99 for any  $s = 0, 1, \dots, k$ , where

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

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

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

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

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

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

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

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

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



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

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

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

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

106 The proof of Lemma 2 is completed.

## 107 2.2 Proof of Lemma 3

108 From (26), we have

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

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

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

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

113 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

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

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

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

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

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

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

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

117 As for the term  $\|h_i(t) - h_i(t-1)\|^2$  in above equation, from (12) and Assumption 1', we have

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

118 The proof of Lemma 3 is completed.

### 119 **3 Proof of Theorem 2**

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

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

122 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  
123 each iteration. We define

$$G(t) = [\nabla f_1(x_1(t), \xi_1(t)), \dots, \nabla f_N(x_N(t), \xi_N(t))],$$

$$\bar{G}(t) = \frac{1}{N} \sum_{i=1}^N \nabla f_i(x_i(t), \xi_i(t)).$$

124 Similarly to (4), we have

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

125 From Assumption 1 and (42), we have

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

126 Then, we have

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

127 Thus, we have

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

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

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

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

130 where  $C_2 = 6\alpha^2 L^2 N B_2 (\alpha^2 \mathbb{E}[\|\bar{\nabla} f(X(0))\|^2] + A_5^2) + N B_1$  and  $A_5 = 2\mathbb{E}[\|X(1)\|^2] +$   
 131  $2\mathbb{E}[\|X(0)\|^2]$ .

132 From (43) and Lemma 5, we have

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

133 If the stepsize  $0 < \alpha \leq \frac{1}{\sqrt{13B_2L}}$ , we have  $B_4 \geq \alpha^2 L^2 B_2$  and  $1 - L\alpha - \frac{6\alpha^4 L^4 B_2}{B_4} > 0$ . Thus, we  
 134 have

$$\begin{aligned} & \frac{1}{T} \sum_{t=1}^T \mathbb{E} \|\nabla f(\bar{X}(t))\|^2 \leq \frac{2}{\alpha T} \left\{ \mathbb{E}[f(\bar{X}(1))] - f(x^*) \right\} \\ & \quad + \frac{C_2 L^2}{B_4 NT} + L\alpha \sigma^2 + \frac{12B_2}{1 - 12\alpha^2 L^2 B_2} \alpha^2 L^2 \sigma^2. \end{aligned} \quad (44)$$

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

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

136 Similar to (35), we have

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

137 In addition, from  $0 < \alpha \leq \frac{1}{\sqrt{13B_2L}}$ , we have

$$1 - 12\alpha^2 B_2 L^2 \geq \frac{1}{13}. \quad (47)$$

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

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

The proof of Theorem 2 is completed.

### 3.1 Proof of Lemma 5

Similar to Lemma 3, we have

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

where

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

In addition, we have

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

From (48) and (49), we have

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

We already have

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

145 In addition, we have

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

146 Thus, from (51), we have

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

147 From (50) and (52), we have

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

148 Thus, from (15), we have

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

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

## 151 References

- 152 [1] Sergio Bittanti and Patrizio Colaneri. Invariant representations of discrete-time periodic systems. *Automatica*,  
153 36(12):1777–1793, 2000.
- 154 [2] Hanlin Tang, Xiangru Lian, Ming Yan, Ce Zhang, and Ji Liu.  $D^2$ : Decentralized training over decentralized  
155 data. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages  
156 4848–4856, 2018.