

# Decentralized Online Optimization

December 27, 2018

## Abstract

ddd

## 1 Problem setup

For any  $i \in [n]$  and  $t \in [T]$ , the random variable  $\xi_{i,t}$  is subject to a distribution  $D_{i,t}$ , that is,

$$\xi_{i,t} \sim D_{i,t}.$$

Besides, a set of random variables  $\Xi_{n,T}$  and the corresponding set of distributions are defined by

$$\Xi_{n,T} = \{\xi_{i,t}\}_{1 \leq i \leq n, 1 \leq t \leq T}, \text{ and } \mathcal{D}_{n,T} = \{D_{i,t}\}_{1 \leq i \leq n, 1 \leq t \leq T},$$

respectively. For math brevity, we use the notation  $\Xi_{n,T} \sim \mathcal{D}_{n,T}$  to represent that  $\xi_{i,t} \sim D_{i,t}$  holds for any  $i \in [n]$  and  $t \in [T]$ .

For any online algorithm  $A \in \mathcal{A}$ , define its dynamic regret as

$$\mathcal{R}_T^A = \mathbb{E}_{\Xi_{n,T} \sim \mathcal{D}_{n,T}} \left( \sum_{i=1}^n \sum_{t=1}^T f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t}) - f_{i,t}(\mathbf{x}_t^*; \xi_{i,t}) \right),$$

where, for any  $\mathbf{x}$ ,

$$f_{i,t}(\mathbf{x}; \xi_{i,t}) := \beta g_{i,t}(\mathbf{x}) + (1 - \beta) h_t(\mathbf{x}; \xi_{i,t})$$

with  $0 < \beta < 1$ , and  $\xi_{i,t}$  is a random variable drawn from an unknown distribution  $D_{i,t}$ .  $g_{i,t}$  is an adversary loss function.  $h_t(\cdot, \xi_{i,t})$  is a given loss function depending on the random variable  $\xi_{i,t}$ . Besides, we denote

$$H_t(\cdot) = \mathbb{E}_{\xi_{i,t} \sim D_{i,t}} h_t(\cdot; \xi_{i,t}),$$

and

$$F_{i,t}(\cdot) = \mathbb{E}_{\xi_{i,t} \sim D_{i,t}} f_{i,t}(\cdot; \xi_{i,t}).$$

The budget of the dynamics is defined as

$$\sum_{t=1}^T \|\mathbf{x}_{t+1}^* - \mathbf{x}_t^*\| \leq M. \tag{1}$$

---

**Algorithm 1** DOG: Decentralized Online Gradient.

---

**Require:** The learning rate  $\eta$ , number of iterations  $T$ , and the confusion matrix  $\mathbf{W}$ .

- 1: **for**  $t = 1, 2, \dots, T$  **do**  
    For the  $i$ -th node with  $i \in [n]$ :
    - 2:     Predict  $\mathbf{x}_{i,t}$ .
    - 3:     Observe the loss function  $f_{i,t}$ ,  
       and suffer loss  $f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t})$ .
  - Update:
    - 4:     Query the gradient  $\nabla f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t})$ .
    - 5:      $\mathbf{x}_{i,t+1} = \sum_{j=1}^n \mathbf{W}_{i,j} \mathbf{x}_{j,t} - \eta \nabla f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t})$ .
- 

## 2 Algorithm

The decentralized online gradient method, namely *DOG*, is presented in Algorithm 1. Comparing with the sequential online gradient method, every node needs to collect the decision variables from its neighbours, and then update its decision variable. The update rule is

$$\mathbf{x}_{i,t+1} = \sum_{j=1}^n \mathbf{W}_{i,j} \mathbf{x}_{j,t} - \eta \nabla f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t}).$$

Here,  $\mathbf{W} \in \mathbb{R}^{n \times n}$  is the confusion matrix. It is a doubly stochastic matrix, which implies that every element of  $\mathbf{W}$  is non-negative,  $\mathbf{W}\mathbf{1} = \mathbf{1}$ , and  $\mathbf{1}^T \mathbf{W} = \mathbf{1}^T$ .

## 3 Theoretical analysis

### 3.1 Assumptions

**Assumption 1.** *We make the following assumptions.*

- For any  $i \in [n]$ ,  $t \in [T]$ , and  $\mathbf{x}$ , there exists a constant  $G$  such that

$$\max \left\{ \mathbb{E}_{\xi_{i,t} \sim D_{i,t}} \|\nabla h_t(\mathbf{x}; \xi_{i,t})\|^2, \|\nabla g_{i,t}(\mathbf{x})\|^2 \right\} \leq G,$$

and

$$\mathbb{E}_{\xi_{i,t} \sim D_{i,t}} \|\nabla h_t(\mathbf{x}; \xi_{i,t}) - \nabla H_t(\mathbf{x})\|^2 \leq \sigma_{i,t}^2.$$

- For any  $\mathbf{x}$  and  $\mathbf{y}$ , we assume  $\|\mathbf{x} - \mathbf{y}\|^2 \leq R$ .
- For any  $i \in [n]$  and  $t \in [T]$ , we assume the function  $f_{i,t}$  is convex and differentiable, and the function  $H_t$  has  $L$ -Lipschitz gradients.

**Assumption 2.** *For any sequence  $\{\mathbf{u}_t\}_{t=2}^T$ , there exists a constant  $V$  such that*

$$\sum_{t=1}^{T-1} (H_{t+1}(\mathbf{u}_{t+1}) - H_t(\mathbf{u}_{t+1})) \leq V.$$

Recall that  $H_t(\cdot) = \mathbb{E}_{\xi_{i,t} \sim D_{i,t}} h_t(\cdot; \xi_{i,t})$ . Assumption 2 implies that the distribution  $D_{i,t}$  cannot changes arbitrarily with the increase of  $t$ .

**Theorem 1.** Denote

$$C_1 = 15G\beta + 4\beta\sigma_{i,t-1}^2 + \frac{6G\eta}{(1-\rho)^2} \left( L + \frac{\eta L^2}{\beta} + 6\eta L^2 + \frac{\beta}{\eta} \right) + 3GL\eta \left( \frac{1}{\beta} + 4 \right);$$

$$C_2 = \left( L + \frac{\eta L^2}{\beta} + 6\eta L^2 + \frac{\beta}{\eta} \right) \frac{4\eta^2}{(1-\rho)^2} + \frac{\beta(\eta + 8nL\eta^2) + 2nL\eta^2}{n\beta}.$$

Using Assumption 1, and choosing  $\eta > 0$  in Algorithm 1, we have

$$\mathcal{R}_T^{DOG} \leq C_1\eta nT + \left( \frac{1}{\beta} + 4 \right) n \mathbb{E}_{\Xi_{n,T} \sim \mathcal{D}_{n,T}} (H_1(\bar{\mathbf{x}}_1) - H_T(\bar{\mathbf{x}}_{T+1}) + V) + C_2 \sum_{t=1}^T \sum_{i=1}^n \sigma_{i,t}^2 + \frac{n}{2\eta} (4\sqrt{RM} + R),$$

holds for any  $0 \leq \beta \leq 1$ .

**Corollary 1.** Using Assumption 1, and choosing

$$\eta = \sqrt{\frac{nM}{T(1+n\beta)}}$$

in Algorithm 1, we have

$$\mathcal{R}_T^{DOG} \lesssim \mathcal{O} \left( \sqrt{(1+n\beta)nMT} \right).$$

## Appendix

**Proof to Theorem 1:**

*Proof.*

$$\begin{aligned} & \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \frac{1}{n} \sum_{i=1}^n f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t}) - f_t(\mathbf{x}_t^*; \xi_{i,t}) \\ &= \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \frac{1}{n} \sum_{i=1}^n \beta (g_{i,t}(\mathbf{x}_{i,t}) - g_{i,t}(\mathbf{x}_t^*)) + (1-\beta) \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \frac{1}{n} \sum_{i=1}^n (h_t(\mathbf{x}_{i,t}; \xi_{i,t}) - h_t(\mathbf{x}_t^*; \xi_{i,t})) \\ &\leq \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \frac{1}{n} \sum_{i=1}^n \beta \langle \nabla g_{i,t}(\mathbf{x}_{i,t}), \mathbf{x}_{i,t} - \mathbf{x}_t^* \rangle + (1-\beta) \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \frac{1}{n} \sum_{i=1}^n \langle \nabla h_t(\mathbf{x}_{i,t}; \xi_{i,t}), \mathbf{x}_{i,t} - \mathbf{x}_t^* \rangle \\ &= \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \frac{1}{n} \sum_{i=1}^n \beta (\langle \nabla g_{i,t}(\mathbf{x}_{i,t}), \mathbf{x}_{i,t} - \bar{\mathbf{x}}_t \rangle + \langle \nabla g_{i,t}(\mathbf{x}_{i,t}), \bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1} \rangle + \langle \nabla g_{i,t}(\mathbf{x}_{i,t}), \bar{\mathbf{x}}_{t+1} - \mathbf{x}_t^* \rangle) \\ &\quad + \frac{1}{n} \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \sum_{i=1}^n (1-\beta) (\langle \nabla h_t(\mathbf{x}_{i,t}; \xi_{i,t}), \mathbf{x}_{i,t} - \bar{\mathbf{x}}_t \rangle + \langle \nabla h_t(\mathbf{x}_{i,t}; \xi_{i,t}), \bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1} \rangle) \\ &\quad + \frac{1}{n} \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \sum_{i=1}^n (1-\beta) (\langle \nabla h_t(\mathbf{x}_{i,t}; \xi_{i,t}), \bar{\mathbf{x}}_{t+1} - \mathbf{x}_t^* \rangle) \\ &= \underbrace{\mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \frac{1}{n} \sum_{i=1}^n \beta (\langle \nabla g_{i,t}(\mathbf{x}_{i,t}), \mathbf{x}_{i,t} - \bar{\mathbf{x}}_t \rangle + \langle \nabla g_{i,t}(\mathbf{x}_{i,t}), \bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1} \rangle)}_{I_1(t)} \end{aligned}$$

$$\begin{aligned}
& + \underbrace{\mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \frac{1}{n} \sum_{i=1}^n (1-\beta) (\langle \nabla h_t(\mathbf{x}_{i,t}; \xi_{i,t}), \mathbf{x}_{i,t} - \bar{\mathbf{x}}_t \rangle + \langle \nabla h_t(\mathbf{x}_{i,t}; \xi_{i,t}), \bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1} \rangle)}_{I_2(t)} \\
& + \underbrace{\mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \left\langle \frac{1}{n} \sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t}), \bar{\mathbf{x}}_{t+1} - \mathbf{x}_t^* \right\rangle}_{I_3(t)}
\end{aligned}$$

Now, we begin to bound  $I_1(t)$ .

$$\begin{aligned}
I_1(t) & \stackrel{\textcircled{1}}{\leq} \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \frac{\beta}{n} \sum_{i=1}^n \left( \frac{\eta}{2} \|\nabla g_{i,t}(\mathbf{x}_{i,t})\|^2 + \frac{1}{2\eta} \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + \frac{\eta}{2} \|\nabla g_{i,t}(\mathbf{x}_{i,t})\|^2 + \frac{1}{2\eta} \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2 \right) \\
& \stackrel{\textcircled{2}}{\leq} \beta G\eta + \frac{\beta}{2n\eta} \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \sum_{i=1}^n \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + \frac{\beta}{2\eta} \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2.
\end{aligned}$$

① holds due to  $\langle \mathbf{a}, \mathbf{b} \rangle \leq \frac{\eta}{2} \|\mathbf{a}\|^2 + \frac{1}{2\eta} \|\mathbf{b}\|^2$  holds for any  $\eta > 0$ . ② holds due to our assumption, that is,  $\|\nabla g_{i,t}(\mathbf{x}_{i,t})\|^2 \leq G$ .

Now, we begin to bound  $I_2(t)$ .

$$I_2(t) = (1-\beta) \left( \underbrace{\mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \frac{1}{n} \sum_{i=1}^n \langle \nabla h_t(\mathbf{x}_{i,t}; \xi_{i,t}), \mathbf{x}_{i,t} - \bar{\mathbf{x}}_t \rangle}_{J_1(t)} + \underbrace{\mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \left\langle \frac{1}{n} \sum_{i=1}^n \nabla h_t(\mathbf{x}_{i,t}; \xi_{i,t}), \bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1} \right\rangle}_{J_2(t)} \right).$$

For  $J_1(t)$ , we have

$$\begin{aligned}
& J_1(t) \\
& = \frac{1}{n} \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \sum_{i=1}^n \langle \nabla h_t(\mathbf{x}_{i,t}; \xi_{i,t}), \mathbf{x}_{i,t} - \bar{\mathbf{x}}_t \rangle \\
& = \frac{1}{n} \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \sum_{i=1}^n \langle \nabla h_t(\mathbf{x}_{i,t}; \xi_{i,t}) - \nabla H_t(\bar{\mathbf{x}}_t), \mathbf{x}_{i,t} - \bar{\mathbf{x}}_t \rangle + \frac{1}{n} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \sum_{i=1}^n \langle \nabla H_t(\bar{\mathbf{x}}_t), \mathbf{x}_{i,t} - \bar{\mathbf{x}}_t \rangle \\
& = \frac{1}{n} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \sum_{i=1}^n \langle \nabla H_t(\mathbf{x}_{i,t}) - \nabla H_t(\bar{\mathbf{x}}_t), \mathbf{x}_{i,t} - \bar{\mathbf{x}}_t \rangle + \frac{1}{n} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \sum_{i=1}^n \langle \nabla H_t(\bar{\mathbf{x}}_t), \mathbf{x}_{i,t} - \bar{\mathbf{x}}_t \rangle \\
& \stackrel{\textcircled{1}}{\leq} \frac{L}{n} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \sum_{i=1}^n \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + \frac{1}{n} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \sum_{i=1}^n \langle \nabla H_t(\bar{\mathbf{x}}_t), \mathbf{x}_{i,t} - \bar{\mathbf{x}}_t \rangle \\
& \stackrel{\textcircled{2}}{\leq} \frac{L}{n} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \sum_{i=1}^n \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + \frac{1}{n} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \sum_{i=1}^n \left( \frac{\eta}{2\nu} \|\nabla H_t(\bar{\mathbf{x}}_t)\|^2 + \frac{\nu}{2\eta} \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 \right) \\
& \leq \frac{L}{n} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \sum_{i=1}^n \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + \frac{\eta}{2\nu} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \|\nabla H_t(\bar{\mathbf{x}}_t)\|^2 + \frac{\nu}{2\eta n} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \sum_{i=1}^n \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2. \quad (2)
\end{aligned}$$

① holds due to  $H_t$  has  $L$ -Lipschitz gradients. ② holds because that  $\langle \mathbf{a}, \mathbf{b} \rangle \leq \frac{\nu}{2} \|\mathbf{a}\|^2 + \frac{1}{2\nu} \|\mathbf{b}\|^2$  holds for any  $\nu > 0$ .

According to Lemma 2, we have

$$\frac{\eta}{2} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \|\nabla H_t(\bar{\mathbf{x}}_t)\|^2$$

$$\begin{aligned}
&\leq \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} (H_t(\bar{\mathbf{x}}_t) - H_t(\bar{\mathbf{x}}_{t+1})) + 2\eta\beta^2(3G + 2\sigma_{i,t-1}^2) + \frac{\eta L^2(1-\beta)^2}{n} \sum_{i=1}^n \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \|\bar{\mathbf{x}}_t - \mathbf{x}_{i,t}\|^2 \\
&\quad + 3GL\eta^2 + \frac{2L\eta^2}{n} \sum_{i=1}^n \sigma_{i,t}^2.
\end{aligned} \tag{3}$$

Substituting (3) into (2), we obtain

$$\begin{aligned}
&J_1(t) \\
&\leq \frac{L}{n} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \sum_{i=1}^n \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + \frac{1}{\nu} \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} (H_t(\bar{\mathbf{x}}_t) - H_t(\bar{\mathbf{x}}_{t+1})) \\
&\quad + \frac{2\eta\beta^2(3G + 2\sigma_{i,t-1}^2)}{\nu} + \frac{\eta L^2(1-\beta)^2}{n\nu} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \sum_{i=1}^n \|\bar{\mathbf{x}}_t - \mathbf{x}_{i,t}\|^2 \\
&\quad + \frac{3GL\eta^2}{\nu} + \frac{2L\eta^2}{n\nu} \sum_{i=1}^n \sigma_{i,t}^2 + \frac{\nu}{2n\eta} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \sum_{i=1}^n \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 \\
&= \left( \frac{L}{n} + \frac{\eta L^2(1-\beta)^2}{n\nu} + \frac{\nu}{2n\eta} \right) \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \sum_{i=1}^n \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + \frac{1}{\nu} \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} (H_t(\bar{\mathbf{x}}_t) - H_t(\bar{\mathbf{x}}_{t+1})) \\
&\quad + \frac{\eta(2\beta^2(3G + 2\sigma_{i,t-1}^2) + 3GL\eta)}{\nu} + \frac{2L\eta^2}{n\nu} \sum_{i=1}^n \sigma_{i,t}^2 \\
&\leq \left( \frac{L}{n} + \frac{\eta L^2}{n\nu} + \frac{\nu}{2n\eta} \right) \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \sum_{i=1}^n \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + \frac{1}{\nu} \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} (H_t(\bar{\mathbf{x}}_t) - H_t(\bar{\mathbf{x}}_{t+1})) \\
&\quad + \frac{\eta(2\beta^2(3G + 2\sigma_{i,t-1}^2) + 3GL\eta)}{\nu} + \frac{2L\eta^2}{n\nu} \sum_{i=1}^n \sigma_{i,t}^2.
\end{aligned}$$

For  $J_2(t)$ , we have

$$\begin{aligned}
&J_2(t) \\
&= \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \left\langle \frac{1}{n} \sum_{i=1}^n \nabla h_t(\mathbf{x}_{i,t}; \xi_{i,t}), \bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1} \right\rangle \\
&\leq \frac{\eta}{2} \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \left\| \frac{1}{n} \sum_{i=1}^n \nabla h_t(\mathbf{x}_{i,t}; \xi_{i,t}) \right\|^2 + \frac{1}{2\eta} \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2 \\
&\leq \frac{\eta}{2} \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \left\| \frac{1}{n} \sum_{i=1}^n (\nabla h_t(\mathbf{x}_{i,t}; \xi_{i,t}) - \nabla H_t(\mathbf{x}_{i,t}) + \nabla H_t(\bar{\mathbf{x}}_t)) \right\|^2 + \frac{1}{2\eta} \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2 \\
&\leq \eta \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \left\| \frac{1}{n} \sum_{i=1}^n (\nabla h_t(\mathbf{x}_{i,t}; \xi_{i,t}) - \nabla H_t(\mathbf{x}_{i,t})) \right\|^2 + \eta \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \left\| \frac{1}{n} \sum_{i=1}^n \nabla H_t(\mathbf{x}_{i,t}) \right\|^2 \\
&\quad + \frac{1}{2\eta} \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2 \\
&\stackrel{\textcircled{1}}{\leq} \frac{\eta}{n^2} \sum_{i=1}^n \sigma_{i,t}^2 + \eta \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \left\| \frac{1}{n} \sum_{i=1}^n (\nabla H_t(\mathbf{x}_{i,t}) - \nabla H_t(\bar{\mathbf{x}}_t) + \nabla H_t(\bar{\mathbf{x}}_t)) \right\|^2 + \frac{1}{2\eta} \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2 \\
&\leq \frac{\eta}{n^2} \sum_{i=1}^n \sigma_{i,t}^2 + 2\eta \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \left\| \frac{1}{n} \sum_{i=1}^n (\nabla H_t(\mathbf{x}_{i,t}) - \nabla H_t(\bar{\mathbf{x}}_t)) \right\|^2
\end{aligned}$$

$$\begin{aligned}
& + 2\eta \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \|\nabla H_t(\bar{\mathbf{x}}_t)\|^2 + \frac{1}{2\eta} \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2 \\
& \stackrel{\textcircled{2}}{\leq} \frac{\eta}{n^2} \sum_{i=1}^n \sigma_{i,t}^2 + \frac{2\eta L^2}{n} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \sum_{i=1}^n \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 \\
& + 2\eta \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \|\nabla H_t(\bar{\mathbf{x}}_t)\|^2 + \frac{1}{2\eta} \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2.
\end{aligned} \tag{4}$$

① holds due to

$$\begin{aligned}
& \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \left\| \frac{1}{n} \sum_{i=1}^n (\nabla h_t(\mathbf{x}_{i,t}; \xi_{i,t}) - \nabla H_t(\mathbf{x}_{i,t})) \right\|^2 \\
& = \frac{1}{n^2} \left( \sum_{i=1}^n \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \|\nabla h_t(\mathbf{x}_{i,t}; \xi_{i,t}) - \nabla H_t(\mathbf{x}_{i,t})\|^2 \right) \\
& + \frac{1}{n^2} \left( 2 \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \sum_{i=1}^n \sum_{j=1, j \neq i}^n \langle \nabla h_t(\mathbf{x}_{i,t}; \xi_{i,t}) - \nabla H_t(\mathbf{x}_{i,t}), \nabla h_t(\mathbf{x}_{j,t}; \xi_{j,t}) - \nabla H_t(\mathbf{x}_{j,t}) \rangle \right) \\
& = \frac{1}{n^2} \sum_{i=1}^n \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \|\nabla h_t(\mathbf{x}_{i,t}; \xi_{i,t}) - \nabla H_t(\mathbf{x}_{i,t})\|^2 + 0 \\
& \leq \frac{1}{n^2} \sum_{i=1}^n \sigma_{i,t}^2.
\end{aligned}$$

② holds due to  $H_t$  has  $L$  Lipschitz gradients.

Substituting (3) into (4), and we have

$$\begin{aligned}
& J_2(t) \\
& \leq \frac{\eta}{n^2} \sum_{i=1}^n \sigma_{i,t}^2 + \frac{2\eta L^2}{n} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \sum_{i=1}^n \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + 4 \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} (H_t(\bar{\mathbf{x}}_t) - H_t(\bar{\mathbf{x}}_{t+1})) + 8G\eta\beta^2 \\
& + \frac{4\eta L^2(1-\beta)^2}{n} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \sum_{i=1}^n \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + 12GL\eta^2 + \frac{8L\eta^2}{n} \sum_{i=1}^n \sigma_{i,t}^2 + \frac{1}{2\eta} \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2 \\
& \leq \frac{\eta + 8nL\eta^2}{n^2} \sum_{i=1}^n \sigma_{i,t}^2 + (2\eta L^2 + 4\eta L^2(1-\beta)^2) \frac{1}{n} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \sum_{i=1}^n \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + 4 \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} (H_t(\bar{\mathbf{x}}_t) - H_t(\bar{\mathbf{x}}_{t+1})) \\
& + 8G\eta\beta^2 + 12GL\eta^2 + \frac{1}{2\eta} \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2 \\
& \leq \frac{\eta + 8nL\eta^2}{n^2} \sum_{i=1}^n \sigma_{i,t}^2 + \frac{6\eta L^2}{n} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \sum_{i=1}^n \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + 4 \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} (H_t(\bar{\mathbf{x}}_t) - H_t(\bar{\mathbf{x}}_{t+1})) \\
& + 8G\eta\beta^2 + 12GL\eta^2 + \frac{1}{2\eta} \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2.
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
& I_2(t) \\
& = (1-\beta)(J_1(t) + J_2(t)) \\
& = (1-\beta) \left( \frac{L}{n} + \frac{\eta L^2}{n\nu} + \frac{\nu}{2n\eta} + \frac{6\eta L^2}{n} \right) \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \sum_{i=1}^n \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + (1-\beta) \left( \frac{1}{\nu} + 4 \right) \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} (H_t(\bar{\mathbf{x}}_t) - H_t(\bar{\mathbf{x}}_{t+1}))
\end{aligned}$$

$$\begin{aligned}
& + (1 - \beta) \left( 2\eta\beta^2 \left( \frac{3G + 2\sigma_{i,t-1}^2}{\nu} + 4 \right) + \frac{1}{2\eta} \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2 \right) \\
& + (1 - \beta) 3LG\eta^2 \left( \frac{1}{\nu} + 4 \right) + (1 - \beta) \frac{\nu(\eta + 8nL\eta^2) + 2nL\eta^2}{n^2\nu} \sum_{i=1}^n \sigma_{i,t}^2 \\
\leq & \left( \frac{L}{n} + \frac{\eta L^2}{n\nu} + \frac{\nu}{2n\eta} + \frac{6\eta L^2}{n} \right) \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \sum_{i=1}^n \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + \left( \frac{1}{\nu} + 4 \right) \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} (H_t(\bar{\mathbf{x}}_t) - H_t(\bar{\mathbf{x}}_{t+1})) \\
& + 2\eta\beta^2 \left( \frac{3G + 2\sigma_{i,t-1}^2}{\nu} + 4 \right) + \frac{1 - \beta}{2\eta} \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2 + 3LG\eta^2 \left( \frac{1}{\nu} + 4 \right) + \frac{\nu(\eta + 8nL\eta^2) + 2nL\eta^2}{n^2\nu} \sum_{i=1}^n \sigma_{i,t}^2.
\end{aligned}$$

Combine those bounds of  $I_1(t)$  and  $I_2(t)$ . We thus have

$$\begin{aligned}
& I_1(t) + I_2(t) \\
\leq & \beta G\eta + \frac{\beta}{2n\eta} \sum_{i=1}^n \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + \frac{\beta}{2\eta} \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2 \\
& + \left( \frac{L}{n} + \frac{\eta L^2}{n\nu} + \frac{\nu}{2n\eta} + \frac{6\eta L^2}{n} \right) \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \sum_{i=1}^n \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + \left( \frac{1}{\nu} + 4 \right) \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} (H_t(\bar{\mathbf{x}}_t) - H_t(\bar{\mathbf{x}}_{t+1})) \\
& + 2\eta\beta^2 \left( \frac{3G + 2\sigma_{i,t-1}^2}{\nu} + 4 \right) + \frac{1 - \beta}{2\eta} \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2 + 3LG\eta^2 \left( \frac{1}{\nu} + 4 \right) + \frac{\nu(\eta + 8nL\eta^2) + 2nL\eta^2}{n^2\nu} \sum_{i=1}^n \sigma_{i,t}^2 \\
\leq & \left( 1 + 2\beta \left( \frac{3}{\nu} + 4 \right) \right) \beta G\eta + \frac{4\eta\beta^2\sigma_{i,t-1}^2}{\nu} + \left( \frac{L}{n} + \frac{\eta L^2}{n\nu} + \frac{\nu}{2n\eta} + \frac{6\eta L^2}{n} + \frac{\beta}{2n\eta} \right) \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \sum_{i=1}^n \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 \\
& + \left( \frac{1}{\nu} + 4 \right) \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} (H_t(\bar{\mathbf{x}}_t) - H_t(\bar{\mathbf{x}}_{t+1})) + \frac{1}{2\eta} \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2 \\
& + 3LG\eta^2 \left( \frac{1}{\nu} + 4 \right) + \frac{\nu(\eta + 8nL\eta^2) + 2nL\eta^2}{n^2\nu} \sum_{i=1}^n \sigma_{i,t}^2.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
& \sum_{t=1}^T (I_1(t) + I_2(t)) \\
\leq & \left( 1 + 2\beta \left( \frac{3}{\nu} + 4 \right) \right) T\beta G\eta + \frac{4T\eta\beta^2\sigma_{i,t-1}^2}{\nu} + \left( \frac{L}{n} + \frac{\eta L^2}{n\nu} + \frac{\nu}{2n\eta} + \frac{6\eta L^2}{n} + \frac{\beta}{2n\eta} \right) \mathbb{E}_{\Xi_{n,T} \sim \mathcal{D}_{n,T}} \sum_{t=1}^T \sum_{i=1}^n \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 \\
& + \left( \frac{1}{\nu} + 4 \right) \mathbb{E}_{\Xi_{n,T} \sim \mathcal{D}_{n,T}} \sum_{t=1}^T (H_t(\bar{\mathbf{x}}_t) - H_t(\bar{\mathbf{x}}_{t+1})) + \frac{1}{2\eta} \mathbb{E}_{\Xi_{n,T} \sim \mathcal{D}_{n,T}} \sum_{t=1}^T \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2 \\
& + 3TLG\eta^2 \left( \frac{1}{\nu} + 4 \right) + \frac{\nu(\eta + 8nL\eta^2) + 2nL\eta^2}{n^2\nu} \sum_{t=1}^T \sum_{i=1}^n \sigma_{i,t}^2. \tag{5}
\end{aligned}$$

According to Assumption 2, we have

$$\begin{aligned}
& \sum_{t=1}^T (H_t(\bar{\mathbf{x}}_t) - H_t(\bar{\mathbf{x}}_{t+1})) \\
= & (H_1(\bar{\mathbf{x}}_1) - H_T(\bar{\mathbf{x}}_{T+1})) + \sum_{t=1}^{T-1} (H_{t+1}(\bar{\mathbf{x}}_{t+1}) - H_t(\bar{\mathbf{x}}_{t+1}))
\end{aligned}$$

$$\leq (H_1(\bar{\mathbf{x}}_1) - H_T(\bar{\mathbf{x}}_{T+1})) + V. \quad (6)$$

Substituting (6) into (5), we obtain

$$\begin{aligned} & \sum_{t=1}^T (I_1(t) + I_2(t)) \\ & \leq \left(1 + 2\beta \left(\frac{3}{\nu} + 4\right)\right) T\beta G\eta + \frac{4T\eta\beta^2\sigma_{i,t-1}^2}{\nu} + \left(\frac{L}{n} + \frac{\eta L^2}{n\nu} + \frac{\nu}{2n\eta} + \frac{6\eta L^2}{n} + \frac{\beta}{2n\eta}\right) \mathbb{E}_{\Xi_{n,T} \sim \mathcal{D}_{n,T}} \sum_{t=1}^T \sum_{i=1}^n \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 \\ & \quad + \left(\frac{1}{\nu} + 4\right) \mathbb{E}_{\Xi_{n,T} \sim \mathcal{D}_{n,T}} (H_1(\bar{\mathbf{x}}_1) - H_T(\bar{\mathbf{x}}_{T+1}) + V) + \frac{1}{2\eta} \mathbb{E}_{\Xi_{n,T} \sim \mathcal{D}_{n,T}} \sum_{t=1}^T \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2 \\ & \quad + 3TLG\eta^2 \left(\frac{1}{\nu} + 4\right) + \frac{\nu(\eta + 8nL\eta^2) + 2nL\eta^2}{n^2\nu} \sum_{t=1}^T \sum_{i=1}^n \sigma_{i,t}^2. \end{aligned}$$

Now, we begin to bound  $I_3(t)$ . Recall that the update rule is

$$\mathbf{x}_{i,t+1} = \sum_{j=1}^n \mathbf{W}_{ij} \mathbf{x}_{j,t} - \eta \nabla f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t}).$$

According to Lemma 3, we have

$$\bar{\mathbf{x}}_{t+1} = \bar{\mathbf{x}}_t - \eta \left( \frac{1}{n} \sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t}) \right). \quad (7)$$

Denote a new auxiliary function  $\phi(\mathbf{z})$  as

$$\phi(\mathbf{z}) = \left\langle \frac{1}{n} \sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t}), \mathbf{z} \right\rangle + \frac{1}{2\eta} \|\mathbf{z} - \bar{\mathbf{x}}_t\|^2.$$

It is trivial to verify that (7) satisfies the first-order optimality condition of the optimization problem:  $\min_{\mathbf{z} \in \mathbb{R}^d} \phi(\mathbf{z})$ , that is,

$$\nabla \phi(\bar{\mathbf{x}}_{t+1}) = \mathbf{0}.$$

We thus have

$$\begin{aligned} \bar{\mathbf{x}}_{t+1} &= \operatorname{argmin}_{\mathbf{z} \in \mathbb{R}^d} \phi(\mathbf{z}) \\ &= \operatorname{argmin}_{\mathbf{z} \in \mathbb{R}^d} \left\langle \frac{1}{n} \sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t}), \mathbf{z} \right\rangle + \frac{1}{2\eta} \|\mathbf{z} - \bar{\mathbf{x}}_t\|^2. \end{aligned}$$

Furthermore, denote a new auxiliary variable  $\bar{\mathbf{x}}_\tau$  as

$$\bar{\mathbf{x}}_\tau = \bar{\mathbf{x}}_{t+1} + \tau (\mathbf{x}_t^* - \bar{\mathbf{x}}_{t+1}),$$

where  $0 < \tau \leq 1$ . According to the optimality of  $\bar{\mathbf{x}}_{t+1}$ , we have

$$\begin{aligned} 0 &\leq \phi(\bar{\mathbf{x}}_\tau) - \phi(\bar{\mathbf{x}}_{t+1}) \\ &= \left\langle \frac{1}{n} \sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t}), \bar{\mathbf{x}}_\tau - \bar{\mathbf{x}}_{t+1} \right\rangle + \frac{1}{2\eta} (\|\bar{\mathbf{x}}_\tau - \bar{\mathbf{x}}_t\|^2 - \|\bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_t\|^2) \end{aligned}$$



$$\begin{aligned}
&= \left\langle \frac{1}{n} \sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t}), \tau (\mathbf{x}_t^* - \bar{\mathbf{x}}_{t+1}) \right\rangle + \frac{1}{2\eta} \left( \|\bar{\mathbf{x}}_{t+1} + \tau (\mathbf{x}_t^* - \bar{\mathbf{x}}_{t+1}) - \bar{\mathbf{x}}_t\|^2 - \|\bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_t\|^2 \right) \\
&= \left\langle \frac{1}{n} \sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t}), \tau (\mathbf{x}_t^* - \bar{\mathbf{x}}_{t+1}) \right\rangle + \frac{1}{2\eta} \left( \|\tau (\mathbf{x}_t^* - \bar{\mathbf{x}}_{t+1})\|^2 + 2 \langle \tau (\mathbf{x}_t^* - \bar{\mathbf{x}}_{t+1}), \bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_t \rangle \right).
\end{aligned}$$

Note that the above inequality holds for any  $0 < \tau \leq 1$ . Divide  $\tau$  on both sides, and we have

$$\begin{aligned}
I_3(t) &= \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \left\langle \frac{1}{n} \sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t}), \bar{\mathbf{x}}_{t+1} - \mathbf{x}_t^* \right\rangle \\
&\leq \frac{1}{2\eta} \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \left( \lim_{\tau \rightarrow 0^+} \tau \|\mathbf{x}_t^* - \bar{\mathbf{x}}_{t+1}\|^2 + 2 \langle \mathbf{x}_t^* - \bar{\mathbf{x}}_{t+1}, \bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_t \rangle \right) \\
&= \frac{1}{\eta} \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \langle \mathbf{x}_t^* - \bar{\mathbf{x}}_{t+1}, \bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_t \rangle \\
&= \frac{1}{2\eta} \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \left( \|\mathbf{x}_t^* - \bar{\mathbf{x}}_t\|^2 - \|\mathbf{x}_t^* - \bar{\mathbf{x}}_{t+1}\|^2 - \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2 \right). \tag{8}
\end{aligned}$$

Besides, we have

$$\begin{aligned}
&\|\mathbf{x}_{t+1}^* - \bar{\mathbf{x}}_{t+1}\|^2 - \|\mathbf{x}_t^* - \bar{\mathbf{x}}_{t+1}\|^2 \\
&= \|\mathbf{x}_{t+1}^*\|^2 - \|\mathbf{x}_t^*\|^2 - 2 \langle \bar{\mathbf{x}}_{t+1}, -\mathbf{x}_t^* + \mathbf{x}_{t+1}^* \rangle \\
&= (\|\mathbf{x}_{t+1}^*\| - \|\mathbf{x}_t^*\|) (\|\mathbf{x}_{t+1}^*\| + \|\mathbf{x}_t^*\|) - 2 \langle \bar{\mathbf{x}}_{t+1}, -\mathbf{x}_t^* + \mathbf{x}_{t+1}^* \rangle \\
&\leq \|\mathbf{x}_{t+1}^* - \mathbf{x}_t^*\| (\|\mathbf{x}_{t+1}^*\| + \|\mathbf{x}_t^*\|) + 2 \|\bar{\mathbf{x}}_{t+1}\| \|\mathbf{x}_{t+1}^* - \mathbf{x}_t^*\| \\
&\leq 4\sqrt{R} \|\mathbf{x}_{t+1}^* - \mathbf{x}_t^*\|.
\end{aligned}$$

The last inequality holds due to our assumption, that is,  $\|\mathbf{x}_{t+1}^*\| = \|\mathbf{x}_{t+1}^* - \mathbf{0}\| \leq \sqrt{R}$ ,  $\|\mathbf{x}_t^*\| = \|\mathbf{x}_t^* - \mathbf{0}\| \leq \sqrt{R}$ , and  $\|\bar{\mathbf{x}}_{t+1}\| = \|\bar{\mathbf{x}}_{t+1} - \mathbf{0}\| \leq \sqrt{R}$ .

Thus, telescoping  $I_3(t)$  over  $t \in [T]$ , we have

$$\begin{aligned}
&\sum_{t=1}^T I_3(t) \\
&\leq \frac{1}{2\eta} \mathbb{E}_{\Xi_{n,T} \sim \mathcal{D}_{n,T}} \left( 4\sqrt{R} \sum_{t=1}^T \|\mathbf{x}_{t+1}^* - \mathbf{x}_t^*\| + \|\bar{\mathbf{x}}_1^* - \bar{\mathbf{x}}_1\|^2 - \|\bar{\mathbf{x}}_T^* - \bar{\mathbf{x}}_{T+1}\|^2 \right) - \frac{1}{2\eta} \mathbb{E}_{\Xi_{n,T} \sim \mathcal{D}_{n,T}} \sum_{t=1}^T \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2 \\
&\leq \frac{1}{2\eta} \left( 4\sqrt{R}M + R \right) - \frac{1}{2\eta} \mathbb{E}_{\Xi_{n,T} \sim \mathcal{D}_{n,T}} \sum_{t=1}^T \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2.
\end{aligned}$$

Here,  $M$  the budget of the dynamics, which is defined in (1).

Combining those bounds of  $I_1(t)$ ,  $I_2(t)$  and  $I_3(t)$  together, we finally obtain

$$\begin{aligned}
&\mathbb{E}_{\Xi_{n,T} \sim \mathcal{D}_{n,T}} \sum_{t=1}^T \sum_{i=1}^n f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t}) - f_t(\mathbf{x}_t^*; \xi_{i,t}) \\
&\leq n \sum_{t=1}^T (I_1(t) + I_2(t) + I_3(t)) \\
&\leq \left( 1 + 2\beta \left( \frac{3}{\nu} + 4 \right) \right) nT\beta G\eta + \frac{4nT\eta\beta^2\sigma_{i,t-1}^2}{\nu} + \left( L + \frac{\eta L^2}{\nu} + \frac{\nu}{2\eta} + 6\eta L^2 + \frac{\beta}{2\eta} \right) \mathbb{E}_{\Xi_{n,T} \sim \mathcal{D}_{n,T}} \sum_{t=1}^T \sum_{i=1}^n \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2
\end{aligned}$$

$$\begin{aligned}
& + \left(\frac{1}{\nu} + 4\right) n \mathbb{E}_{\Xi_{n,T} \sim \mathcal{D}_{n,T}} (H_1(\bar{\mathbf{x}}_1) - H_T(\bar{\mathbf{x}}_{T+1}) + V) \\
& + 3nTLG\eta^2 \left(\frac{1}{\nu} + 4\right) + \frac{\nu(\eta + 8nL\eta^2) + 2nL\eta^2}{n\nu} \sum_{t=1}^T \sum_{i=1}^n \sigma_{i,t}^2 + \frac{n}{2\eta} (4\sqrt{RM} + R) \\
& \stackrel{\textcircled{1}}{\leq} \left(1 + 2\beta \left(\frac{3}{\nu} + 4\right)\right) nT\beta G\eta + \frac{4nT\eta\beta^2\sigma_{i,t-1}^2}{\nu} + \left(L + \frac{\eta L^2}{\nu} + \frac{\nu}{2\eta} + 6\eta L^2 + \frac{\beta}{2\eta}\right) \frac{\eta^2}{(1-\rho)^2} \left(6nTG + 4 \sum_{t=1}^T \sum_{i=1}^n \sigma_{i,t}^2\right) \\
& + \left(\frac{1}{\nu} + 4\right) n \mathbb{E}_{\Xi_{n,T} \sim \mathcal{D}_{n,T}} (H_1(\bar{\mathbf{x}}_1) - H_T(\bar{\mathbf{x}}_{T+1}) + V) \\
& + 3nTLG\eta^2 \left(\frac{1}{\nu} + 4\right) + \frac{\nu(\eta + 8nL\eta^2) + 2nL\eta^2}{n\nu} \sum_{t=1}^T \sum_{i=1}^n \sigma_{i,t}^2 + \frac{n}{2\eta} (4\sqrt{RM} + R).
\end{aligned}$$

① holds due to Lemma 4, that is,

$$\mathbb{E}_{\Xi_{n,T} \sim \mathcal{D}_{n,T}} \sum_{i=1}^n \sum_{t=1}^T \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 \leq \frac{\eta^2}{(1-\rho)^2} \left(6nTG + 4 \sum_{t=1}^T \sum_{i=1}^n \sigma_{i,t}^2\right).$$

When  $\nu = \beta$ , we finally have

$$\begin{aligned}
& \mathcal{R}_T^{DOG} \\
& = \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \sum_{t=1}^T \sum_{i=1}^n f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t}) - f_t(\mathbf{x}_t^*; \xi_{i,t}) \\
& \leq 15nT\beta G\eta + 4nT\eta\beta\sigma_{i,t-1}^2 + \left(L + \frac{\eta L^2}{\beta} + 6\eta L^2 + \frac{\beta}{\eta}\right) \frac{\eta^2}{(1-\rho)^2} \left(6nTG + 4 \sum_{t=1}^T \sum_{i=1}^n \sigma_{i,t}^2\right) \\
& + \left(\frac{1}{\beta} + 4\right) n \mathbb{E}_{\Xi_{n,T} \sim \mathcal{D}_{n,T}} (H_1(\bar{\mathbf{x}}_1) - H_T(\bar{\mathbf{x}}_{T+1}) + V) + 3nTLG\eta^2 \left(\frac{1}{\beta} + 4\right) \\
& + \frac{\beta(\eta + 8nL\eta^2) + 2nL\eta^2}{n\beta} \sum_{t=1}^T \sum_{i=1}^n \sigma_{i,t}^2 + \frac{n}{2\eta} (4\sqrt{RM} + R) \\
& = \eta nT \left(15G\beta + 4\beta\sigma_{i,t-1}^2 + \frac{6G\eta}{(1-\rho)^2} \left(L + \frac{\eta L^2}{\beta} + 6\eta L^2 + \frac{\beta}{\eta}\right) + 3GL\eta \left(\frac{1}{\beta} + 4\right)\right) \\
& + \left(\frac{1}{\beta} + 4\right) n \mathbb{E}_{\Xi_{n,T} \sim \mathcal{D}_{n,T}} (H_1(\bar{\mathbf{x}}_1) - H_T(\bar{\mathbf{x}}_{T+1}) + V) \\
& + \left(\left(L + \frac{\eta L^2}{\beta} + 6\eta L^2 + \frac{\beta}{\eta}\right) \frac{4\eta^2}{(1-\rho)^2} + \frac{\beta(\eta + 8nL\eta^2) + 2nL\eta^2}{n\beta}\right) \sum_{t=1}^T \sum_{i=1}^n \sigma_{i,t}^2 + \frac{n}{2\eta} (4\sqrt{RM} + R).
\end{aligned}$$

It completes the proof.  $\square$

**Lemma 1.** Using Assumption 1, we have

$$\mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \|\nabla f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t})\|^2 \leq 6G + 4\sigma_{i,t}^2.$$

*Proof.*

$$\mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \|\nabla f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t})\|^2$$

$$\begin{aligned}
&= \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \|\beta \nabla g_{i,t}(\mathbf{x}_{i,t}) + (1-\beta) \nabla h_t(\mathbf{x}_{i,t}; \xi_{i,t})\|^2 \\
&\leq 2\beta^2 \|\nabla g_{i,t}(\mathbf{x}_{i,t})\|^2 + 2(1-\beta)^2 \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \|\nabla h_t(\mathbf{x}_{i,t}; \xi_{i,t})\|^2 \\
&\leq 2G\beta^2 + 2(1-\beta)^2 \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \|\nabla h_t(\mathbf{x}_{i,t}; \xi_{i,t}) - \nabla H_t(\mathbf{x}_{i,t}) + \nabla H_t(\mathbf{x}_{i,t})\|^2 \\
&\leq 2G\beta^2 + 4(1-\beta)^2 \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \|\nabla h_t(\mathbf{x}_{i,t}; \xi_{i,t}) - \nabla H_t(\mathbf{x}_{i,t})\|^2 + 4(1-\beta)^2 \|\nabla H_t(\mathbf{x}_{i,t})\|^2 \\
&\leq 2G\beta^2 + 4(1-\beta)^2 \sigma_{i,t}^2 + 4(1-\beta)^2 G \\
&\leq 6G + 4\sigma_{i,t}^2.
\end{aligned}$$

The last inequality holds due to  $0 \leq \beta \leq 1$ .  $\square$

**Lemma 2.** Using Assumption 1, and setting  $\eta > 0$  in Algorithm 1, we have

$$\begin{aligned}
&\frac{\eta}{2} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \|\nabla H_t(\bar{\mathbf{x}}_t)\|^2 \\
&\leq \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} (H_t(\bar{\mathbf{x}}_t) - H_t(\bar{\mathbf{x}}_{t+1})) + 2\eta\beta^2(3G + 2\sigma_{i,t-1}^2) + \frac{\eta L^2(1-\beta)^2}{n} \sum_{i=1}^n \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \|\bar{\mathbf{x}}_t - \mathbf{x}_{i,t}\|^2 \\
&\quad + 3GL\eta^2 + \frac{2L\eta^2}{n} \sum_{i=1}^n \sigma_{i,t}^2.
\end{aligned} \tag{9}$$

*Proof.*

$$\begin{aligned}
&\mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} H_t(\bar{\mathbf{x}}_{t+1}) \\
&\leq \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} H_t(\bar{\mathbf{x}}_t) + \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \langle \nabla H_t(\bar{\mathbf{x}}_t), \bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_t \rangle + \frac{L}{2} \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \|\bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_t\|^2 \\
&= \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} H_t(\bar{\mathbf{x}}_t) + \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \left\langle \nabla H_t(\bar{\mathbf{x}}_t), -\frac{\eta}{n} \sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t}) \right\rangle + \frac{L}{2} \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \left\| \frac{\eta}{n} \sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t}) \right\|^2 \\
&\stackrel{\textcircled{1}}{\leq} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} H_t(\bar{\mathbf{x}}_t) + \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \left\langle \nabla H_t(\bar{\mathbf{x}}_t), -\frac{\eta}{n} \sum_{i=1}^n \nabla F_{i,t}(\mathbf{x}_{i,t}) \right\rangle + \frac{L\eta^2}{2n} \sum_{i=1}^n \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \|\nabla f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t})\|^2 \\
&= \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} H_t(\bar{\mathbf{x}}_t) + \frac{\eta}{2} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \left( \left\| \nabla H_t(\bar{\mathbf{x}}_t) - \frac{1}{n} \sum_{i=1}^n \nabla F_{i,t}(\mathbf{x}_{i,t}) \right\|^2 - \|\nabla H_t(\bar{\mathbf{x}}_t)\|^2 - \left\| \frac{1}{n} \sum_{i=1}^n \nabla F_{i,t}(\mathbf{x}_{i,t}) \right\|^2 \right) \\
&\quad + \frac{L\eta^2}{2n} \sum_{i=1}^n \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \|\nabla f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t})\|^2 \\
&\stackrel{\textcircled{2}}{\leq} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} H_t(\bar{\mathbf{x}}_t) + \frac{\eta}{2} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \left\| \nabla H_t(\bar{\mathbf{x}}_t) - \frac{1}{n} \sum_{i=1}^n (\beta \nabla g_{i,t}(\mathbf{x}_{i,t}) + (1-\beta) \nabla H_t(\mathbf{x}_{i,t})) \right\|^2 \\
&\quad - \frac{\eta}{2} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \|\nabla H_t(\bar{\mathbf{x}}_t)\|^2 - \frac{\eta}{2} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \left\| \frac{1}{n} \sum_{i=1}^n \nabla F_{i,t}(\mathbf{x}_{i,t}) \right\|^2 + 3GL\eta^2 + \frac{2L\eta^2}{n} \sum_{i=1}^n \sigma_{i,t}^2 \\
&\leq \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} H_t(\bar{\mathbf{x}}_t) - \frac{\eta}{2} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \|\nabla H_t(\bar{\mathbf{x}}_t)\|^2 - \frac{\eta}{2} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \left\| \frac{1}{n} \sum_{i=1}^n \nabla F_{i,t}(\mathbf{x}_{i,t}) \right\|^2 + 3GL\eta^2 + \frac{2L\eta^2}{n} \sum_{i=1}^n \sigma_{i,t}^2 \\
&\quad + \frac{\eta}{2} \left( 2\beta^2 \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \left\| \nabla H_t(\bar{\mathbf{x}}_t) - \frac{1}{n} \sum_{i=1}^n \nabla g_{i,t}(\mathbf{x}_{i,t}) \right\|^2 + 2(1-\beta)^2 \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \left\| \nabla H_t(\bar{\mathbf{x}}_t) - \frac{1}{n} \sum_{i=1}^n \nabla H_t(\mathbf{x}_{i,t}) \right\|^2 \right)
\end{aligned}$$

$$\begin{aligned}
&\leq \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} H_t(\bar{\mathbf{x}}_t) - \frac{\eta}{2} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \|\nabla H_t(\bar{\mathbf{x}}_t)\|^2 - \frac{\eta}{2} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \left\| \frac{1}{n} \sum_{i=1}^n \nabla F_{i,t}(\mathbf{x}_{i,t}) \right\|^2 + 3GL\eta^2 + \frac{2L\eta^2}{n} \sum_{i=1}^n \sigma_{i,t}^2 \\
&\quad + \frac{\eta}{2n} \sum_{i=1}^n \left( 2\beta^2 \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \|\nabla H_t(\bar{\mathbf{x}}_t) - \nabla g_{i,t}(\mathbf{x}_{i,t})\|^2 + 2(1-\beta)^2 \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \|\nabla H_t(\bar{\mathbf{x}}_t) - \nabla H_t(\mathbf{x}_{i,t})\|^2 \right) \\
&\leq \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} H_t(\bar{\mathbf{x}}_t) - \frac{\eta}{2} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \|\nabla H_t(\bar{\mathbf{x}}_t)\|^2 - \frac{\eta}{2} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \left\| \frac{1}{n} \sum_{i=1}^n \nabla F_{i,t}(\mathbf{x}_{i,t}) \right\|^2 + 3GL\eta^2 + \frac{2L\eta^2}{n} \sum_{i=1}^n \sigma_{i,t}^2 \\
&\quad + \frac{\eta}{2n} \sum_{i=1}^n \left( 2\beta^2 \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \left( 2\|\nabla H_t(\bar{\mathbf{x}}_t)\|^2 + 2\|\nabla g_{i,t}(\mathbf{x}_{i,t})\|^2 \right) + 2(1-\beta)^2 \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \|\nabla H_t(\bar{\mathbf{x}}_t) - \nabla H_t(\mathbf{x}_{i,t})\|^2 \right) \\
&\stackrel{\textcircled{3}}{\leq} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} H_t(\bar{\mathbf{x}}_t) - \frac{\eta}{2} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \|\nabla H_t(\bar{\mathbf{x}}_t)\|^2 - \frac{\eta}{2} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \left\| \frac{1}{n} \sum_{i=1}^n \nabla F_{i,t}(\mathbf{x}_{i,t}) \right\|^2 + 3GL\eta^2 + \frac{2L\eta^2}{n} \sum_{i=1}^n \sigma_{i,t}^2 \\
&\quad + \frac{\eta}{2n} \sum_{i=1}^n \left( 2\beta^2 \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \left( 2\|\nabla H_t(\bar{\mathbf{x}}_t)\|^2 + 2G \right) + 2(1-\beta)^2 L^2 \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \|\bar{\mathbf{x}}_t - \mathbf{x}_{i,t}\|^2 \right) \\
&\stackrel{\textcircled{4}}{\leq} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} H_t(\bar{\mathbf{x}}_t) + \frac{\eta}{2} \left( 4\beta^2(3G + 2\sigma_{i,t-1}^2) + 2(1-\beta)^2 L^2 \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \|\bar{\mathbf{x}}_t - \mathbf{x}_{i,t}\|^2 \right) \\
&\quad - \frac{\eta}{2} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \|\nabla H_t(\bar{\mathbf{x}}_t)\|^2 - \frac{\eta}{2} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \left\| \frac{1}{n} \sum_{i=1}^n \nabla F_{i,t}(\mathbf{x}_{i,t}) \right\|^2 + 3GL\eta^2 + \frac{2L\eta^2}{n} \sum_{i=1}^n \sigma_{i,t}^2 \\
&= \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} H_t(\bar{\mathbf{x}}_t) + 2\eta\beta^2(3G + 2\sigma_{i,t-1}^2) + \eta(1-\beta)^2 L^2 \left( \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \|\bar{\mathbf{x}}_t - \mathbf{x}_{i,t}\|^2 \right) \\
&\quad - \frac{\eta}{2} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \|\nabla H_t(\bar{\mathbf{x}}_t)\|^2 - \frac{\eta}{2} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \left\| \frac{1}{n} \sum_{i=1}^n \nabla F_{i,t}(\mathbf{x}_{i,t}) \right\|^2 + 3GL\eta^2 + \frac{2L\eta^2}{n} \sum_{i=1}^n \sigma_{i,t}^2.
\end{aligned}$$

① holds due to  $F_{i,t}(\mathbf{x}_{i,t}) = \mathbb{E}_{\xi_{i,t} \sim D_{i,t}} f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t})$ , and  $\left\| \frac{1}{n} \sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t}) \right\|^2 \leq \frac{1}{n} \sum_{i=1}^n \|\nabla f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t})\|^2$ .

② holds due to Lemma 1. ③ holds due to our assumption, that is,  $\|\nabla g_{i,t}(\mathbf{x}_{i,t})\|^2 \leq G$ , and  $H_t$  has  $L$  Lipschitz gradient. ④ holds due to

$$\begin{aligned}
&\mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \|\nabla H_t(\bar{\mathbf{x}}_t)\|^2 \\
&= \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \|\nabla H_t(\bar{\mathbf{x}}_t) - \nabla h_t(\bar{\mathbf{x}}_t; \xi_{i,t-1}) + \nabla h_t(\bar{\mathbf{x}}_t; \xi_{i,t-1})\|^2 \\
&\leq 2 \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \|\nabla H_t(\bar{\mathbf{x}}_t) - \nabla h_t(\bar{\mathbf{x}}_t; \xi_{i,t-1})\|^2 + 2 \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \|\nabla h_t(\bar{\mathbf{x}}_t; \xi_{i,t-1})\|^2 \\
&\leq 2\sigma_{i,t-1}^2 + 2G.
\end{aligned}$$

Finally, we obtain

$$\begin{aligned}
&\frac{\eta}{2} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \|\nabla H_t(\bar{\mathbf{x}}_t)\|^2 \\
&\leq \frac{\eta}{2} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \|\nabla H_t(\bar{\mathbf{x}}_t)\|^2 + \frac{\eta}{2} \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \left\| \frac{1}{n} \sum_{i=1}^n \nabla F_{i,t}(\mathbf{x}_{i,t}) \right\|^2 \\
&\leq \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} (H_t(\bar{\mathbf{x}}_t) - H_t(\bar{\mathbf{x}}_{t+1})) + 2\eta\beta^2(3G + 2\sigma_{i,t-1}^2) + \frac{\eta L^2(1-\beta)^2}{n} \sum_{i=1}^n \mathbb{E}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \|\bar{\mathbf{x}}_t - \mathbf{x}_{i,t}\|^2
\end{aligned}$$

$$+ 3GL\eta^2 + \frac{2L\eta^2}{n} \sum_{i=1}^n \sigma_{i,t}^2.$$

It completes the proof.  $\square$

**Lemma 3.** Denote  $\bar{\mathbf{x}}_t = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_{i,t}$ . We have

$$\bar{\mathbf{x}}_{t+1} = \bar{\mathbf{x}}_t - \eta \left( \frac{1}{n} \sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t}) \right).$$

*Proof.* Denote

$$\begin{aligned} \mathbf{X}_t &= [\mathbf{x}_{1,t}, \mathbf{x}_{2,t}, \dots, \mathbf{x}_{n,t}] \in \mathbb{R}^{d \times n}, \\ \mathbf{G}_t &= [\nabla f_{1,t}(\mathbf{x}_{1,t}; \xi_{1,t}), \nabla f_{2,t}(\mathbf{x}_{2,t}; \xi_{2,t}), \dots, \nabla f_{n,t}(\mathbf{x}_{n,t}; \xi_{n,t})] \in \mathbb{R}^{d \times n}. \end{aligned}$$

Recall that

$$\mathbf{x}_{i,t+1} = \sum_{j=1}^n \mathbf{W}_{ij} \mathbf{x}_{j,t} - \eta \nabla f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t}).$$

Equivalently, we re-formulate the update rule as

$$\mathbf{X}_{t+1} = \mathbf{X}_t \mathbf{W} - \eta \mathbf{G}_t.$$

Since the confusion matrix  $\mathbf{W}$  is doubly stochastic, we have

$$\mathbf{W} \mathbf{1} = \mathbf{1}.$$

Thus, we have

$$\begin{aligned} \bar{\mathbf{x}}_{t+1} &= \frac{1}{n} \sum_{i=1}^n \mathbf{x}_{i,t+1} \\ &= \mathbf{X}_{t+1} \frac{\mathbf{1}}{n} \\ &= \mathbf{X}_t \mathbf{W} \frac{\mathbf{1}}{n} - \eta \mathbf{G}_t \frac{\mathbf{1}}{n} \\ &= \mathbf{X}_t \frac{\mathbf{1}}{n} - \eta \mathbf{G}_t \frac{\mathbf{1}}{n} \\ &= \bar{\mathbf{x}}_t - \eta \left( \frac{1}{n} \sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t}) \right). \end{aligned}$$

$\square$

**Lemma 4.** Using Assumption 1, and setting  $\eta > 0$  in Algorithm 1, we have

$$\mathbb{E}_{\Xi_{n,T} \sim \mathcal{D}_{n,T}} \sum_{i=1}^n \sum_{t=1}^T \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 \leq \frac{\eta^2}{(1-\rho)^2} \left( 6nTG + 4 \sum_{t=1}^T \sum_{i=1}^n \sigma_{i,t}^2 \right).$$

*Proof.* Recall that

$$\mathbf{x}_{i,t+1} = \sum_{j=1}^n \mathbf{W}_{ij} \mathbf{x}_{j,t} - \eta \nabla f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t}),$$

and according to Lemma 3, we have

$$\bar{\mathbf{x}}_{t+1} = \bar{\mathbf{x}}_t - \eta \left( \frac{1}{n} \sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t}) \right).$$

Denote

$$\begin{aligned} \mathbf{X}_t &= [\mathbf{x}_{1,t}, \mathbf{x}_{2,t}, \dots, \mathbf{x}_{n,t}] \in \mathbb{R}^{d \times n}, \\ \mathbf{G}_t &= [\nabla f_{1,t}(\mathbf{x}_{1,t}; \xi_{1,t}), \nabla f_{2,t}(\mathbf{x}_{2,t}; \xi_{2,t}), \dots, \nabla f_{n,t}(\mathbf{x}_{n,t}; \xi_{n,t})] \in \mathbb{R}^{d \times n}. \end{aligned}$$

By letting  $\mathbf{x}_{i,1} = \mathbf{0}$  for any  $i \in [n]$ , the update rule is re-formulated as

$$\mathbf{X}_{t+1} = \mathbf{X}_t \mathbf{W} - \eta \mathbf{G}_t = - \sum_{s=1}^t \eta \mathbf{G}_s \mathbf{W}^{t-s}.$$

Similarly, denote  $\bar{\mathbf{G}}_t = \frac{1}{n} \sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t})$ , and we have

$$\bar{\mathbf{x}}_{t+1} = \bar{\mathbf{x}}_t - \eta \left( \frac{1}{n} \sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t}) \right) = - \sum_{s=1}^t \eta \bar{\mathbf{G}}_s. \quad (10)$$

Therefore,

$$\begin{aligned} & \sum_{i=1}^n \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 \\ & \stackrel{\textcircled{1}}{=} \sum_{i=1}^n \left\| \sum_{s=1}^{t-1} \eta \bar{\mathbf{G}}_s - \eta \mathbf{G}_s \mathbf{W}^{t-s-1} \mathbf{e}_i \right\|^2 \\ & \stackrel{\textcircled{2}}{=} \left\| \sum_{s=1}^{t-1} \eta \mathbf{G}_s \mathbf{v}_1 \mathbf{v}_1^T - \eta \mathbf{G}_s \mathbf{W}^{t-s-1} \right\|_F^2 \\ & \stackrel{\textcircled{3}}{\leq} \left( \eta \rho^{t-s-1} \left\| \sum_{s=1}^{t-1} \mathbf{G}_s \right\|_F \right)^2 \\ & \leq \left( \sum_{s=1}^{t-1} \eta \rho^{t-s-1} \|\mathbf{G}_s\|_F \right)^2. \end{aligned}$$

① holds due to  $\mathbf{e}_i$  is a unit basis vector, whose  $i$ -th element is 1 and other elements are 0s. ② holds due to  $\mathbf{v}_1 = \frac{1_n}{\sqrt{n}}$ . ③ holds due to Lemma 5.

Thus, we have

$$\begin{aligned} & \mathbb{E}_{\Xi_{n,T} \sim \mathcal{D}_{n,T}} \sum_{i=1}^n \sum_{t=1}^T \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 \\ & \leq \mathbb{E}_{\Xi_{n,T} \sim \mathcal{D}_{n,T}} \sum_{t=1}^T \left( \sum_{s=1}^{t-1} \eta \rho^{t-s-1} \|\mathbf{G}_s\|_F \right)^2 \\ & \stackrel{\textcircled{1}}{\leq} \frac{\eta^2}{(1-\rho)^2} \mathbb{E}_{\Xi_{n,T} \sim \mathcal{D}_{n,T}} \left( \sum_{t=1}^T \|\mathbf{G}_t\|_F^2 \right) \\ & = \frac{\eta^2}{(1-\rho)^2} \left( \mathbb{E}_{\Xi_{n,T} \sim \mathcal{D}_{n,T}} \sum_{t=1}^T \sum_{i=1}^n \|\nabla f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t})\|^2 \right) \end{aligned}$$

$$\begin{aligned}
&= \frac{\eta^2}{(1-\rho)^2} \left( \sum_{t=1}^T \sum_{i=1}^n \mathbb{E}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \|\nabla f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t})\|^2 \right) \\
&\stackrel{\textcircled{2}}{\leq} \frac{\eta^2}{(1-\rho)^2} \left( 6nTG + 4 \sum_{t=1}^T \sum_{i=1}^n \sigma_{i,t}^2 \right).
\end{aligned}$$

① holds due to Lemma 6. ② holds due to Lemma 1. It completes the proof.  $\square$

**Lemma 5** (Appeared in Lemma 5 in [Tang et al., 2018]). *For any matrix  $\mathbf{X}_t \in \mathbb{R}^{d \times n}$ , decompose the confusion matrix  $\mathbf{W}$  as  $\mathbf{W} = \sum_{i=1}^n \lambda_i \mathbf{v}_i \mathbf{v}_i^\top = \mathbf{P} \mathbf{\Lambda} \mathbf{P}^\top$ , where  $\mathbf{P} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n] \in \mathbb{R}^{n \times n}$ ,  $\mathbf{v}_i$  is the normalized eigenvector of  $\lambda_i$ .  $\mathbf{\Lambda}$  is a diagonal matrix, and  $\lambda_i$  be its  $i$ -th element. We have*

$$\|\mathbf{X}_t \mathbf{W}^t - \mathbf{X}_t \mathbf{v}_1 \mathbf{v}_1^\top\|_F^2 \leq \|\rho^t \mathbf{X}_t\|_F^2,$$

where  $\rho = \max\{|\lambda_2(\mathbf{W})|, |\lambda_n(\mathbf{W})|\}$ .

**Lemma 6** (Appeared in Lemma 6 in [Tang et al., 2018]). *Given two non-negative sequences  $\{a_t\}_{t=1}^\infty$  and  $\{b_t\}_{t=1}^\infty$  that satisfying*

$$a_t = \sum_{s=1}^t \rho^{t-s} b_s,$$

with  $\rho \in [0, 1)$ , we have

$$\sum_{t=1}^k a_t^2 \leq \frac{1}{(1-\rho)^2} \sum_{s=1}^k b_s^2.$$

## References

H. Tang, S. Gan, C. Zhang, T. Zhang, and J. Liu. Communication Compression for Decentralized Training. *arXiv.org*, Mar. 2018.