Gossip Online Learning: Exchanging Local Models to Track Dynamics

January 5, 2019

Abstract

1 Application scenarios

Communication efficient online learning. Suppose we want to conduct online learning on multiple nodes. Those nodes are connected in a decentralized way. At every iteration, the local model has to be broadcast to its neighbours. During online learning, the communication efficiency also needs to be considered. In the case, $g_{i,t}(\mathbf{x}_{i,t})$ represents the loss incurred by the learning model, and $h_t(\mathbf{x}_{i,t}; \xi_{i,t})$ represents the loss incurred by some a random method, e.g., random quantization method [**?**], to guarantee the communication efficiency.

Online learning with privacy. Suppose we want to conduct online learning on a decentralized network. But, there is a hacker who can sniff at the network, and obtains the transmitted data packages. To protect the privacy, we use a randomization method to handle the local model before transmitting it in the network. In the case, $g_{i,t}(\mathbf{x}_{i,t})$ represents the loss incurred by the learning model. $h_t(\mathbf{x}_{i,t}; \xi_{i,t})$ represents the loss incurred by some a randomization encryption method, e.g., parameter perturbation [**?**], to protect the privacy.

Online music recommendation with unreliable features. Suppose we want to decide whether to recommend a music to Bob by using a public dataset consisting of historical browser records on Youtube. But, some values of features in those records are not reliable. For example, Alice's browser record is in the public dataset. But Alice does not want to let others know her real birthday and age. She submits random numbers for such information when signing up as an Youtube user. Note that those unreliable values, e.g., Alice's age and birthday, usually do not change, which is modeled by an unknown distribution. But, other reliable values, e.g., Alice's perference to a music, may change over time, which is a classic setting for an online learning problem. In the case, $g_{i,t}(\mathbf{x}_{i,t})$ represents the loss incurred by those reliable features in the learing model, e.g., perference to a music. $h_t(\mathbf{x}_{i,t}; \xi_{i,t})$ represents the loss incurred by those unreliable features in the learing model, e.g., age and birthday.

2 Notations

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 $t \in [T]$.

3 Problem setup

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

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

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

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

and

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

$\{\mathbf{x}_t^*\}_{t=1}^{T}$ is the sequence of reference points, and

$$\{\mathbf{x}_t^*\}_{t=1}^{T} \in \left\{ \{\mathbf{z}_t\}_{t=1}^{T} : \sum_{t=1}^{T-1} \|\mathbf{z}_t - \mathbf{z}_{t+1}\| \leq M \right\}.$$

Here, M is the budget of the dynamics, that is,

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

4 Algorithm

Algorithm 1 DOG: Decentralized Online Gradient.

Require: The learning rate η, number of iterations T, and the confusion matrix $\mathbf{W}$. $\mathbf{x}_{i,1} = \mathbf{0}$ for any $i \in [n]$.
 1: **for** $t = 1, 2, ..., T$ **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 a sub-gradient $\partial f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t})$.
 5:       $\mathbf{x}_{i,t+1} = \sum_{j=1}^{n} \mathbf{W}_{i,j} \mathbf{x}_{j,t} - \eta \partial f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t})$.

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

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

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, $\mathbf{W1} = \mathbf{1}$, and $\mathbf{1}^{\mathrm{T}} \mathbf{W} = \mathbf{1}^{\mathrm{T}}$.

5 Theoretical analysis

5.1 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*

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

and

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

- *For any $\mathbf{x}$ and $\mathbf{y}$, we assume $\|\mathbf{x} - \mathbf{y}\|^2 \le R$.*

- *For any $i \in [n]$ and $t \in [T]$, we assume the function $f_{i,t}$ is convex, but may be non-smooth. Furthermore, we assume the function H_t has L-Lipschitz gradients. In a nutshell, $g_{i,t}$ may be non-convex, non-smooth. H_t is smooth, but may be non-convex. $f_{i,t}$ is convex, but may be non-smooth.*

Theorem 1. *Denote*

$$C_0 := \frac{1}{\sqrt{\beta^2 + \eta}} + 4;$$

$$C_1 := \frac{\beta}{2\eta} + L + \frac{\sqrt{\beta^2 + \eta}}{2\eta} + 2\eta L^2 + C_0(1 - \beta)^2 L^2 \eta.$$

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

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

$$\le \eta T \left(n\beta G + (1 - \beta)\sigma^2 \right) + n(1 - \beta)C_0 \left(\mathop{\mathbb{E}}_{\Xi_{n,T} \sim \mathcal{D}_{n,T}} \sum_{t=1}^{T} (H_t(\bar{\mathbf{x}}_t) - H_t(\bar{\mathbf{x}}_{t+1})) \right)$$

$$+ (1 - \beta)\frac{nT\eta^2 G C_1}{(1 - \rho)^2} + n(1 - \beta)C_0 \left(4T\beta^2 \eta G + \frac{TGL\eta^2}{2} \right) + \frac{n}{2\eta} \left(4\sqrt{R}M + R \right).$$

Corollary 1. *Recall that*

$$C_0 = \frac{1}{\sqrt{\beta^2 + \eta}} + 4.$$

Using Assumption 1, and choosing

$$\eta = \sqrt{\frac{nM}{T \left(n\beta G + (1 - \beta)\sigma^2 \right)}}$$

in Algorithm 1, we have

$$\mathcal{R}_T^{DOG} \lesssim \sqrt{nMT \left(\beta nG + (1 - \beta)\sigma^2 \right)} + n(1 - \beta)C_0 \mathop{\mathbb{E}}_{\Xi_{n,T} \sim \mathcal{D}_{n,T}} \sum_{t=1}^{T} (H_t(\bar{\mathbf{x}}_t) - H_t(\bar{\mathbf{x}}_{t+1})).$$

3

Appendix

Proof to Theorem 1:

Proof.

$$\mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \frac{1}{n}\sum_{i=1}^{n} f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t}) - f_{i,t}(\mathbf{x}_t^*;\xi_{i,t})$$

$$\leq \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \frac{1}{n}\sum_{i=1}^{n} \langle \partial f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t}), \mathbf{x}_{i,t} - \mathbf{x}_t^* \rangle$$

$$= \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \frac{1}{n}\sum_{i=1}^{n} \beta \langle \partial g_{i,t}(\mathbf{x}_{i,t}), \mathbf{x}_{i,t} - \mathbf{x}_t^* \rangle + (1-\beta) \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \frac{1}{n}\sum_{i=1}^{n} \langle \nabla h_t(\mathbf{x}_{i,t};\xi_{i,t}), \mathbf{x}_{i,t} - \mathbf{x}_t^* \rangle$$

$$= \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \frac{1}{n}\sum_{i=1}^{n} \beta \left(\langle \partial g_{i,t}(\mathbf{x}_{i,t}), \mathbf{x}_{i,t} - \bar{\mathbf{x}}_t \rangle + \langle \partial g_{i,t}(\mathbf{x}_{i,t}), \bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1} \rangle + \langle \partial g_{i,t}(\mathbf{x}_{i,t}), \bar{\mathbf{x}}_{t+1} - \mathbf{x}_t^* \rangle \right)$$

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

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

$$= \underbrace{\mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \frac{1}{n}\sum_{i=1}^{n} \beta \left(\langle \partial g_{i,t}(\mathbf{x}_{i,t}), \mathbf{x}_{i,t} - \bar{\mathbf{x}}_t \rangle + \langle \partial g_{i,t}(\mathbf{x}_{i,t}), \bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1} \rangle \right)}_{I_1(t)}$$

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

$$+ \underbrace{\mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \left\langle \frac{1}{n}\sum_{i=1}^{n} \partial f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t}), \bar{\mathbf{x}}_{t+1} - \mathbf{x}_t^* \right\rangle}_{I_3(t)}$$

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

$$I_1(t) \overset{\text{①}}{\leq} \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \frac{\beta}{n}\sum_{i=1}^{n} \left(\frac{\eta}{2} \|\partial g_{i,t}(\mathbf{x}_{i,t})\|^2 + \frac{1}{2\eta} \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + \frac{\eta}{2} \|\partial g_{i,t}(\mathbf{x}_{i,t})\|^2 + \frac{1}{2\eta} \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2 \right)$$

$$\leq \beta G\eta + \frac{\beta}{2n\eta} \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \sum_{i=1}^{n} \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + \frac{\beta}{2\eta} \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2 .$$

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 $\eta > 0$.

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

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

For $J_1(t)$, we have

$$J_1(t)$$

$$= \frac{1}{n} \mathop{\mathbb{E}}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \sum_{i=1}^{n} \langle \nabla h_t(\mathbf{x}_{i,t}; \xi_{i,t}), \mathbf{x}_{i,t} - \bar{\mathbf{x}}_t \rangle$$

$$= \frac{1}{n} \mathop{\mathbb{E}}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \sum_{i=1}^{n} \langle \nabla h_t(\mathbf{x}_{i,t}; \xi_{i,t}) - \nabla H_t(\bar{\mathbf{x}}_t), \mathbf{x}_{i,t} - \bar{\mathbf{x}}_t \rangle + \frac{1}{n} \mathop{\mathbb{E}}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \sum_{i=1}^{n} \langle \nabla H_t(\bar{\mathbf{x}}_t), \mathbf{x}_{i,t} - \bar{\mathbf{x}}_t \rangle$$

$$= \frac{1}{n} \mathop{\mathbb{E}}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \sum_{i=1}^{n} \langle \nabla H_t(\mathbf{x}_{i,t}) - \nabla H_t(\bar{\mathbf{x}}_t), \mathbf{x}_{i,t} - \bar{\mathbf{x}}_t \rangle + \frac{1}{n} \mathop{\mathbb{E}}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \sum_{i=1}^{n} \langle \nabla H_t(\bar{\mathbf{x}}_t), \mathbf{x}_{i,t} - \bar{\mathbf{x}}_t \rangle$$

$$\overset{①}{\leq} \frac{L}{n} \mathop{\mathbb{E}}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \sum_{i=1}^{n} \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + \frac{1}{n} \mathop{\mathbb{E}}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \sum_{i=1}^{n} \langle \nabla H_t(\bar{\mathbf{x}}_t), \mathbf{x}_{i,t} - \bar{\mathbf{x}}_t \rangle$$

$$\overset{②}{\leq} \frac{L}{n} \mathop{\mathbb{E}}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \sum_{i=1}^{n} \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + \frac{1}{n} \mathop{\mathbb{E}}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \sum_{i=1}^{n} \left(\frac{\eta}{2\nu} \|\nabla H_t(\bar{\mathbf{x}}_t)\|^2 + \frac{\nu}{2\eta} \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 \right)$$

$$\leq \frac{L}{n} \mathop{\mathbb{E}}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \sum_{i=1}^{n} \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + \frac{\eta}{2\nu} \mathop{\mathbb{E}}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \|\nabla H_t(\bar{\mathbf{x}}_t)\|^2 + \frac{\nu}{2\eta n} \mathop{\mathbb{E}}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \sum_{i=1}^{n} \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 . \quad (2)$$

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 $\nu > 0$.

For $J_2(t)$, we have

$$J_2(t)$$

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

$$\leq \frac{\eta}{2} \mathop{\mathbb{E}}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla h_t(\mathbf{x}_{i,t}; \xi_{i,t}) \right\|^2 + \frac{1}{2\eta} \mathop{\mathbb{E}}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2$$

$$\leq \frac{\eta}{2} \mathop{\mathbb{E}}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \left\| \frac{1}{n} \sum_{i=1}^{n} (\nabla h_t(\mathbf{x}_{i,t}; \xi_{i,t}) - \nabla H_t(\mathbf{x}_{i,t}) + \nabla H_t(\mathbf{x}_{i,t})) \right\|^2 + \frac{1}{2\eta} \mathop{\mathbb{E}}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2$$

$$\leq \eta \mathop{\mathbb{E}}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \left\| \frac{1}{n} \sum_{i=1}^{n} (\nabla h_t(\mathbf{x}_{i,t}; \xi_{i,t}) - \nabla H_t(\mathbf{x}_{i,t})) \right\|^2 + \eta \mathop{\mathbb{E}}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \left\| \frac{1}{n} \sum_{i=1}^{n} \nabla H_t(\mathbf{x}_{i,t}) \right\|^2$$

$$\quad + \frac{1}{2\eta} \mathop{\mathbb{E}}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2$$

$$\overset{①}{\leq} \frac{\eta}{n} \sigma^2 + \eta \mathop{\mathbb{E}}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \left\| \frac{1}{n} \sum_{i=1}^{n} (\nabla H_t(\mathbf{x}_{i,t}) - \nabla H_t(\bar{\mathbf{x}}_t) + \nabla H_t(\bar{\mathbf{x}}_t)) \right\|^2 + \frac{1}{2\eta} \mathop{\mathbb{E}}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2$$

$$\leq \frac{\eta}{n} \sigma^2 + 2\eta \mathop{\mathbb{E}}_{\Xi_{n,t-1} \sim \mathcal{D}_{n,t-1}} \left\| \frac{1}{n} \sum_{i=1}^{n} (\nabla H_t(\mathbf{x}_{i,t}) - \nabla H_t(\bar{\mathbf{x}}_t)) \right\|^2$$

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

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

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

$$\overset{\textcircled{2}}{\leq} \frac{\eta}{n}\sigma^2 + \frac{2\eta L^2}{n} \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}} \sum_{i=1}^{n} \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + 2\eta \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}} \|\nabla H_t(\bar{\mathbf{x}}_t)\|^2 + \frac{1}{2\eta} \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2.$$

① holds due to

$$\mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \left\| \frac{1}{n} \sum_{i=1}^{n} \left(\nabla h_t(\mathbf{x}_{i,t};\xi_{i,t}) - \nabla H_t(\mathbf{x}_{i,t})\right) \right\|^2$$

$$= \frac{1}{n^2} \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}} \left(\sum_{i=1}^{n} \mathop{\mathbb{E}}_{\xi_{i,t}\sim D_{i,t}} \|\nabla h_t(\mathbf{x}_{i,t};\xi_{i,t}) - \nabla H_t(\mathbf{x}_{i,t})\|^2 \right)$$

$$+ \frac{1}{n^2} \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}} \left(2\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} \left\langle \mathop{\mathbb{E}}_{\xi_{i,t}\sim D_{i,t}} \nabla h_t(\mathbf{x}_{i,t};\xi_{i,t}) - \nabla H_t(\mathbf{x}_{i,t}), \mathop{\mathbb{E}}_{\xi_{j,t}\sim D_{j,t}} \nabla h_t(\mathbf{x}_{j,t};\xi_{j,t}) - \nabla H_t(\mathbf{x}_{j,t}) \right\rangle \right)$$

$$= \frac{1}{n^2} \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}} \sum_{i=1}^{n} \mathop{\mathbb{E}}_{\xi_{i,t}\sim D_{i,t}} \|\nabla h_t(\mathbf{x}_{i,t};\xi_{i,t}) - \nabla H_t(\mathbf{x}_{i,t})\|^2 + 0$$

$$\leq \frac{1}{n}\sigma^2.$$

② holds due to H_t has L Lipschitz gradients.

Therefore, we obtain

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

$$+ (1-\beta)\left(\frac{\eta}{n}\sigma^2 + \frac{2\eta L^2}{n} \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}} \sum_{i=1}^{n} \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 \right)$$

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

$$\leq (1-\beta)\left(\frac{L}{n} + \frac{\nu}{2n\eta} + \frac{2\eta L^2}{n} \right) \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}} \sum_{i=1}^{n} \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + \left(\frac{\eta}{2\nu} + 2\eta \right)(1-\beta) \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}} \|\nabla H_t(\bar{\mathbf{x}}_t)\|^2$$

$$+ \frac{\eta(1-\beta)\sigma^2}{n} + \frac{1-\beta}{2\eta} \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2.$$

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

$$I_1(t) + I_2(t)$$
$$\leq \beta G\eta + \frac{\beta}{2n\eta} \sum_{i=1}^{n} \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}} \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + \frac{\beta}{2\eta} \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2$$

$$+ (1-\beta)\left(\frac{L}{n} + \frac{\nu}{2n\eta} + \frac{2\eta L^2}{n} \right) \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}} \sum_{i=1}^{n} \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + \left(\frac{\eta}{2\nu} + 2\eta \right)(1-\beta) \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}} \|\nabla H_t(\bar{\mathbf{x}}_t)\|^2$$

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

$$= \eta\left(\beta G + \frac{(1-\beta)\sigma^2}{n} \right) + (1-\beta)\left(\frac{\beta}{2n\eta} + \frac{L}{n} + \frac{\nu}{2n\eta} + \frac{2\eta L^2}{n} \right) \sum_{i=1}^{n} \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}} \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2$$

6

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

Therefore, we have

$$\sum_{t=1}^{T} (I_1(t) + I_2(t))$$

$$\leq \eta T \left(\beta G + \frac{(1-\beta)\sigma^2}{n} \right) + (1 - \beta) \left(\frac{\beta}{2n\eta} + \frac{L}{n} + \frac{\nu}{2n\eta} + \frac{2\eta L^2}{n} \right) \underset{\Xi_{n,T-1} \sim \mathcal{D}_{n,T-1}}{\mathbb{E}} \sum_{i=1}^{n} \sum_{t=1}^{T} \left\| \mathbf{x}_{i,t} - \bar{\mathbf{x}}_t \right\|^2$$

$$+ \frac{1}{2\eta} \underset{\Xi_{n,T} \sim \mathcal{D}_{n,T}}{\mathbb{E}} \sum_{t=1}^{T} \left\| \bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1} \right\|^2 + \left(\frac{\eta}{2\nu} + 2\eta \right) (1 - \beta) \underset{\Xi_{n,T-1} \sim \mathcal{D}_{n,T-1}}{\mathbb{E}} \sum_{t=1}^{T} \left\| \nabla H_t(\bar{\mathbf{x}}_t) \right\|^2 .$$

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

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

According to Lemma 3, we have

$$\bar{\mathbf{x}}_{t+1} = \bar{\mathbf{x}}_t - \eta \left(\frac{1}{n} \sum_{i=1}^{n} \partial f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t}) \right). \tag{3}$$

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

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

It is trivial to verify that (3) 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

$$\bar{\mathbf{x}}_{t+1} = \underset{\mathbf{z} \in \mathbb{R}^d}{\arg\min} \, \phi(\mathbf{z})$$

$$= \underset{\mathbf{z} \in \mathbb{R}^d}{\arg\min} \left\langle \frac{1}{n} \sum_{i=1}^{n} \partial f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t}), \mathbf{z} \right\rangle + \frac{1}{2\eta} \left\| \mathbf{z} - \bar{\mathbf{x}}_t \right\|^2 .$$

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

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

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

$$0 \leq \phi(\bar{\mathbf{x}}_\tau) - \phi(\bar{\mathbf{x}}_{t+1})$$

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

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

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

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

$$
\begin{aligned}
I_3(t) &= \mathop{\mathbb{E}}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \left\langle \frac{1}{n} \sum_{i=1}^{n} \partial f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t}), \bar{\mathbf{x}}_{t+1} - \mathbf{x}_t^* \right\rangle \\
&\leq \frac{1}{2\eta} \mathop{\mathbb{E}}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \left(\lim_{\tau \to 0^+} \tau \left\| (\mathbf{x}_t^* - \bar{\mathbf{x}}_{t+1}) \right\|^2 + 2\left\langle \mathbf{x}_t^* - \bar{\mathbf{x}}_{t+1}, \bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_t \right\rangle \right) \\
&= \frac{1}{\eta} \mathop{\mathbb{E}}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \left\langle \mathbf{x}_t^* - \bar{\mathbf{x}}_{t+1}, \bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_t \right\rangle \\
&= \frac{1}{2\eta} \mathop{\mathbb{E}}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \left(\left\| \mathbf{x}_t^* - \bar{\mathbf{x}}_t \right\|^2 - \left\| \mathbf{x}_t^* - \bar{\mathbf{x}}_{t+1} \right\|^2 - \left\| \bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1} \right\|^2 \right).
\end{aligned}
\tag{4}
$$

Besides, we have

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

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

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

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

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

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

8

$$\overset{\text{\textcircled{1}}}{\leq} \eta T \left(n\beta G + (1-\beta)\sigma^2 \right) + n(1-\beta)\left(\frac{1}{\nu} + 4 \right) \left(\underset{\Xi_{n,T} \sim \mathcal{D}_{n,T}}{\mathbb{E}} \sum_{t=1}^{T} (H_t(\bar{\mathbf{x}}_t) - H_t(\bar{\mathbf{x}}_{t+1})) \right)$$

$$+ (1-\beta) \left(\frac{\beta}{2\eta} + L + \frac{\nu}{2\eta} + 2\eta L^2 + \left(\frac{1}{\nu} + 4 \right)(1-\beta)^2 L^2 \eta \right) \underset{\Xi_{n,T} \sim \mathcal{D}_{n,T}}{\mathbb{E}} \sum_{t=1}^{T} \sum_{i=1}^{n} \|\bar{\mathbf{x}}_t - \mathbf{x}_{i,t}\|^2$$

$$+ n(1-\beta)\left(\frac{1}{\nu} + 4 \right) \left(4T\beta^2 \eta G + \frac{TGL\eta^2}{2} \right) + \frac{n}{2\eta}\left(4\sqrt{R}M + R \right)$$

$$\overset{\text{\textcircled{2}}}{\leq} \eta T \left(n\beta G + (1-\beta)\sigma^2 \right) + n(1-\beta)\left(\frac{1}{\nu} + 4 \right) \left(\underset{\Xi_{n,T} \sim \mathcal{D}_{n,T}}{\mathbb{E}} \sum_{t=1}^{T} (H_t(\bar{\mathbf{x}}_t) - H_t(\bar{\mathbf{x}}_{t+1})) \right)$$

$$+ (1-\beta) \left(\frac{\beta}{2\eta} + L + \frac{\nu}{2\eta} + 2\eta L^2 + \left(\frac{1}{\nu} + 4 \right)(1-\beta)^2 L^2 \eta \right) \frac{nT\eta^2 G}{(1-\rho)^2}$$

$$+ n(1-\beta)\left(\frac{1}{\nu} + 4 \right) \left(4T\beta^2 \eta G + \frac{TGL\eta^2}{2} \right) + \frac{n}{2\eta}\left(4\sqrt{R}M + R \right).$$

\textcircled{1} holds due to Lemma 2. That is, we have

$$\frac{\eta}{2} \underset{\Xi_{n,T-1} \sim \mathcal{D}_{n,T-1}}{\mathbb{E}} \sum_{t=1}^{T} \|\nabla H_t(\bar{\mathbf{x}}_t)\|^2 \tag{5}$$

$$\leq \underset{\Xi_{n,T} \sim \mathcal{D}_{n,T}}{\mathbb{E}} \sum_{t=1}^{T} (H_t(\bar{\mathbf{x}}_t) - H_t(\bar{\mathbf{x}}_{t+1})) + 4T\beta^2 \eta G + \frac{(1-\beta)^2 L^2 \eta}{n} \underset{\Xi_{n,T-1} \sim \mathcal{D}_{n,T-1}}{\mathbb{E}} \sum_{t=1}^{T} \sum_{i=1}^{n} \|\bar{\mathbf{x}}_t - \mathbf{x}_{i,t}\|^2 + \frac{TGL\eta^2}{2}.$$

\textcircled{2} holds due to Lemma 4

$$\underset{\Xi_{n,T-1} \sim \mathcal{D}_{n,T-1}}{\mathbb{E}} \sum_{i=1}^{n} \sum_{t=1}^{T} \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 \leq \frac{nT\eta^2 G}{(1-\rho)^2}.$$

Letting $\nu = \sqrt{\beta^2 + \eta}$, we have

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

$$\leq \eta T \left(n\beta G + (1-\beta)\sigma^2 \right) + n(1-\beta)\left(\frac{1}{\sqrt{\beta^2 + \eta}} + 4 \right) \left(\underset{\Xi_{n,T} \sim \mathcal{D}_{n,T}}{\mathbb{E}} \sum_{t=1}^{T} (H_t(\bar{\mathbf{x}}_t) - H_t(\bar{\mathbf{x}}_{t+1})) \right)$$

$$+ (1-\beta) \left(\frac{\beta}{2\eta} + L + \frac{\sqrt{\beta^2 + \eta}}{2\eta} + 2\eta L^2 + \left(\frac{1}{\sqrt{\beta^2 + \eta}} + 4 \right)(1-\beta)^2 L^2 \eta \right) \frac{nT\eta^2 G}{(1-\rho)^2}$$

$$+ n(1-\beta)\left(\frac{1}{\sqrt{\beta^2 + \eta}} + 4 \right) \left(4T\beta^2 \eta G + \frac{TGL\eta^2}{2} \right) + \frac{n}{2\eta}\left(4\sqrt{R}M + R \right).$$

It completes the proof.

$\square$

Lemma 1. *Using Assumption 1, we have*

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

Proof.

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

$$= \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \left\| \beta \partial g_{i,t}(\mathbf{x}_{i,t}) + (1-\beta)\nabla h_t(\mathbf{x}_{i,t};\xi_{i,t}) \right\|^2$$

$$\leq \beta \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}} \left\| \partial g_{i,t}(\mathbf{x}_{i,t}) \right\|^2 + (1-\beta) \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \left\| \nabla h_t(\mathbf{x}_{i,t};\xi_{i,t}) \right\|^2$$

$$\leq G.$$

It completes the proof. $\qquad\square$

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

$$\frac{\eta}{2} \mathop{\mathbb{E}}_{\Xi_{n,T-1}\sim\mathcal{D}_{n,T-1}} \sum_{t=1}^{T} \left\| \nabla H_t(\bar{\mathbf{x}}_t) \right\|^2 \tag{6}$$

$$\leq \mathop{\mathbb{E}}_{\Xi_{n,T}\sim\mathcal{D}_{n,T}} \sum_{t=1}^{T} \left(H_t(\bar{\mathbf{x}}_t) - H_t(\bar{\mathbf{x}}_{t+1}) \right) + 4T\beta^2\eta G + \frac{(1-\beta)^2 L^2 \eta}{n} \mathop{\mathbb{E}}_{\Xi_{n,T-1}\sim\mathcal{D}_{n,T-1}} \sum_{t=1}^{T}\sum_{i=1}^{n} \left\| \bar{\mathbf{x}}_t - \mathbf{x}_{i,t} \right\|^2 + \frac{TGL\eta^2}{2}.$$

Proof.

$$\mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} H_t(\bar{\mathbf{x}}_{t+1})$$

$$\leq \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}} H_t(\bar{\mathbf{x}}_t) + \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \left\langle \nabla H_t(\bar{\mathbf{x}}_t), \bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_t \right\rangle + \frac{L}{2} \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \left\| \bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_t \right\|^2$$

$$= \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}} H_t(\bar{\mathbf{x}}_t) + \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \left\langle \nabla H_t(\bar{\mathbf{x}}_t), -\frac{\eta}{n}\sum_{i=1}^{n}\partial f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t}) \right\rangle + \frac{L}{2} \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \left\| \frac{\eta}{n}\sum_{i=1}^{n}\partial f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t}) \right\|^2$$

$$= \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}} H_t(\bar{\mathbf{x}}_t) + \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}} \left\langle \nabla H_t(\bar{\mathbf{x}}_t), -\frac{\eta}{n}\sum_{i=1}^{n}\partial f_{i,t}(\mathbf{x}_{i,t}) \right\rangle + \frac{L}{2} \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \left\| \frac{\eta}{n}\sum_{i=1}^{n}\partial f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t}) \right\|^2. \tag{7}$$

Besides, we have

$$\mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}} \left\langle \nabla H_t(\bar{\mathbf{x}}_t), -\frac{\eta}{n}\sum_{i=1}^{n}\partial f_{i,t}(\mathbf{x}_{i,t}) \right\rangle$$

$$= \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}} \frac{\eta}{2} \left(\left\| \nabla H_t(\bar{\mathbf{x}}_t) - \frac{1}{n}\sum_{i=1}^{n}\partial f_{i,t}(\mathbf{x}_{i,t}) \right\|^2 - \left\| \nabla H_t(\bar{\mathbf{x}}_t) \right\|^2 - \left\| \frac{1}{n}\sum_{i=1}^{n}\partial f_{i,t}(\mathbf{x}_{i,t}) \right\|^2 \right)$$

$$\leq \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}} \frac{\eta}{2} \left(\left\| \nabla H_t(\bar{\mathbf{x}}_t) - \frac{1}{n}\sum_{i=1}^{n}\left(\beta\partial g_{i,t}(\mathbf{x}_{i,t}) + (1-\beta)\nabla H_t(\mathbf{x}_{i,t})\right) \right\|^2 \right) - \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}} \frac{\eta}{2} \left\| \nabla H_t(\bar{\mathbf{x}}_t) \right\|^2$$

$$\leq \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}} \frac{\eta}{2} \left(2\beta^2 \left\| \nabla H_t(\bar{\mathbf{x}}_t) - \frac{1}{n}\sum_{i=1}^{n}\partial g_{i,t}(\mathbf{x}_{i,t}) \right\|^2 + 2(1-\beta)^2 \left\| \nabla H_t(\bar{\mathbf{x}}_t) - \frac{1}{n}\sum_{i=1}^{n}\nabla H_t(\mathbf{x}_{i,t}) \right\|^2 \right)$$

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

$$\leq \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}} \frac{\eta}{2} \left(2\beta^2 \left\| \nabla H_t(\bar{\mathbf{x}}_t) - \frac{1}{n}\sum_{i=1}^{n}\partial g_{i,t}(\mathbf{x}_{i,t}) \right\|^2 + \frac{2(1-\beta)^2}{n}\sum_{i=1}^{n} \left\| \nabla H_t(\bar{\mathbf{x}}_t) - \nabla H_t(\mathbf{x}_{i,t}) \right\|^2 \right)$$

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

10

$$\leq \underset{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}{\mathbb{E}}\frac{\eta}{2}\left(2\beta^2\left\|\nabla H_t(\bar{\mathbf{x}}_t)-\frac{1}{n}\sum_{i=1}^n\partial g_{i,t}(\mathbf{x}_{i,t})\right\|^2+\frac{2(1-\beta)^2L^2}{n}\sum_{i=1}^n\|\bar{\mathbf{x}}_t-\mathbf{x}_{i,t}\|^2\right)-\underset{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}{\mathbb{E}}\frac{\eta}{2}\|\nabla H_t(\bar{\mathbf{x}}_t)\|^2$$

$$\leq \underset{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}{\mathbb{E}}\frac{\eta}{2}\left(4\beta^2\|\nabla H_t(\bar{\mathbf{x}}_t)\|^2+4\beta^2\left\|\frac{1}{n}\sum_{i=1}^n\partial g_{i,t}(\mathbf{x}_{i,t})\right\|^2+\frac{2(1-\beta)^2L^2}{n}\sum_{i=1}^n\|\bar{\mathbf{x}}_t-\mathbf{x}_{i,t}\|^2\right)$$

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

$$\overset{①}{\leq}\underset{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}{\mathbb{E}}\frac{\eta}{2}\left(8\beta^2G+\frac{2(1-\beta)^2L^2}{n}\sum_{i=1}^n\|\bar{\mathbf{x}}_t-\mathbf{x}_{i,t}\|^2\right)-\underset{\Xi_{n,t}\sim\mathcal{D}_{n,t}}{\mathbb{E}}\frac{\eta}{2}\|\nabla H_t(\bar{\mathbf{x}}_t)\|^2. \tag{8}$$

① holds due to

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

$$=\underset{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}{\mathbb{E}}\left\|\underset{\xi_{i,t}\sim D_{i,t}}{\mathbb{E}}\nabla h_t(\bar{\mathbf{x}}_t;\xi_{i,t})\right\|^2$$

$$\leq\underset{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}{\mathbb{E}}\left(\underset{\xi_{i,t}\sim D_{i,t}}{\mathbb{E}}\|\nabla h_t(\bar{\mathbf{x}}_t;\xi_{i,t})\|^2\right),\quad\forall i\in[n]$$

$$\leq G,$$

and

$$\underset{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}{\mathbb{E}}\left\|\frac{1}{n}\sum_{i=1}^n\partial g_{i,t}(\mathbf{x}_{i,t})\right\|^2\leq\frac{1}{n}\sum_{i=1}^n\underset{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}{\mathbb{E}}\|\partial g_{i,t}(\mathbf{x}_{i,t})\|^2\leq G.$$

According to Lemma 1, we have

$$\underset{\Xi_{n,t}\sim\mathcal{D}_{n,t}}{\mathbb{E}}\|\partial f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t})\|^2\leq G. \tag{9}$$

Substituting (8) and (9) into (7), and telescoping $t\in[T]$, we obtain

$$\underset{\Xi_{n,T}\sim\mathcal{D}_{n,T}}{\mathbb{E}}\sum_{t=1}^T H_t(\bar{\mathbf{x}}_{t+1})$$

$$\leq\underset{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}{\mathbb{E}}H_t(\bar{\mathbf{x}}_t)+\underset{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}{\mathbb{E}}\left\langle\nabla H_t(\bar{\mathbf{x}}_t),-\frac{\eta}{n}\sum_{i=1}^n\partial f_{i,t}(\mathbf{x}_{i,t})\right\rangle+\frac{L}{2}\underset{\Xi_{n,t}\sim\mathcal{D}_{n,t}}{\mathbb{E}}\left\|\frac{\eta}{n}\sum_{i=1}^n\partial f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t})\right\|^2$$

$$\leq\underset{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}{\mathbb{E}}H_t(\bar{\mathbf{x}}_t)+\left(\underset{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}{\mathbb{E}}\frac{\eta}{2}\left(8\beta^2G+\frac{2(1-\beta)^2L^2}{n}\sum_{i=1}^n\|\bar{\mathbf{x}}_t-\mathbf{x}_{i,t}\|^2\right)-\underset{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}{\mathbb{E}}\frac{\eta}{2}\|\nabla H_t(\bar{\mathbf{x}}_t)\|^2\right)+\frac{GL\eta^2}{2}$$

$$=\underset{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}{\mathbb{E}}H_t(\bar{\mathbf{x}}_t)+\left(4\eta\beta^2G+\frac{(1-\beta)^2L^2\eta}{n}\underset{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}{\mathbb{E}}\sum_{i=1}^n\|\bar{\mathbf{x}}_t-\mathbf{x}_{i,t}\|^2-\underset{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}{\mathbb{E}}\frac{\eta}{2}\|\nabla H_t(\bar{\mathbf{x}}_t)\|^2\right)+\frac{GL\eta^2}{2}.$$

Telescoping over $t\in[T]$, we have

$$\frac{\eta}{2}\underset{\Xi_{n,T-1}\sim\mathcal{D}_{n,T-1}}{\mathbb{E}}\sum_{t=1}^T\|\nabla H_t(\bar{\mathbf{x}}_t)\|^2 \tag{10}$$

$$\leq\underset{\Xi_{n,T}\sim\mathcal{D}_{n,T}}{\mathbb{E}}\sum_{t=1}^T(H_t(\bar{\mathbf{x}}_t)-H_t(\bar{\mathbf{x}}_{t+1}))+4T\beta^2\eta G+\frac{(1-\beta)^2L^2\eta}{n}\underset{\Xi_{n,T-1}\sim\mathcal{D}_{n,T-1}}{\mathbb{E}}\sum_{t=1}^T\sum_{i=1}^n\|\bar{\mathbf{x}}_t-\mathbf{x}_{i,t}\|^2+\frac{TGL\eta^2}{2}.$$

It completes the proof. $\qquad\square$

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

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

Proof. Denote

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

Recall that

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

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

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

It completes the proof. $\square$

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

$$\mathbb{E}_{\Xi_{n,T} \sim \mathcal{D}_{n,T}} \sum_{i=1}^{n} \sum_{t=1}^{T} \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 \leq \frac{nT\eta^2 G}{(1-\rho)^2}.$$

Proof. Recall that

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

and according to Lemma 3, we have

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

12

Denote

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

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

$$\bar{\mathbf{x}}_{t+1} = \bar{\mathbf{x}}_t - \eta \left(\frac{1}{n} \sum_{i=1}^{n} \partial f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t}) \right) = - \sum_{s=1}^{t} \eta \bar{\mathbf{G}}_s. \tag{11}$$

Therefore,

$$\sum_{i=1}^{n} \| \mathbf{x}_{i,t} - \bar{\mathbf{x}}_t \|^2$$

$$\overset{\text{①}}{=} \sum_{i=1}^{n} \left\| \sum_{s=1}^{t-1} \eta \bar{\mathbf{G}}_s - \eta \mathbf{G}_s \mathbf{W}^{t-s-1} \mathbf{e}_i \right\|^2$$

$$\overset{\text{②}}{=} \left\| \sum_{s=1}^{t-1} \eta \mathbf{G}_s \mathbf{v}_1 \mathbf{v}_1^{\mathrm{T}} - \eta \mathbf{G}_s \mathbf{W}^{t-s-1} \right\|_F^2$$

$$\overset{\text{③}}{\leq} \left(\eta \rho^{t-s-1} \left\| \sum_{s=1}^{t-1} \mathbf{G}_s \right\|_F \right)^2$$

$$\leq \left(\sum_{s=1}^{t-1} \eta \rho^{t-s-1} \| \mathbf{G}_s \|_F \right)^2.$$

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 $\mathbf{v}_1 = \frac{\mathbf{1}_n}{\sqrt{n}}$. ③ holds due to Lemma 5.

Thus, we have

$$\mathbb{E}_{\Xi_{n,T} \sim \mathcal{D}_{n,T}} \sum_{i=1}^{n} \sum_{t=1}^{T} \| \mathbf{x}_{i,t} - \bar{\mathbf{x}}_t \|^2$$

$$\leq \mathbb{E}_{\Xi_{n,T} \sim \mathcal{D}_{n,T}} \sum_{t=1}^{T} \left(\sum_{s=1}^{t-1} \eta \rho^{t-s-1} \| \mathbf{G}_s \|_F \right)^2$$

$$\overset{\text{①}}{\leq} \frac{\eta^2}{(1-\rho)^2} \mathbb{E}_{\Xi_{n,T} \sim \mathcal{D}_{n,T}} \left(\sum_{t=1}^{T} \| \mathbf{G}_t \|_F^2 \right)$$

$$= \frac{\eta^2}{(1-\rho)^2} \left(\mathbb{E}_{\Xi_{n,T} \sim \mathcal{D}_{n,T}} \sum_{t=1}^{T} \sum_{i=1}^{n} \| \partial f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t}) \|^2 \right)$$

$$\overset{\text{②}}{=} \frac{n T \eta^2 G}{(1-\rho)^2}.$$

① holds due to Lemma 6. ② holds due to Lemma 1.

$\square$

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 $\mathbf{W} = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^{\mathrm{T}} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{\mathrm{T}}$, where $\mathbf{P} = [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_n] \in \mathbb{R}^{n \times n}$, $\mathbf{v}_i$ is the normalized eigenvector of λ_i. $\mathbf{\Lambda}$ is a diagonal matrix, and λ_i be its i-th element. We have*

$$\left\| \mathbf{X}_t \mathbf{W}^t - \mathbf{X}_t \mathbf{v}_1 \mathbf{v}_1^{\mathrm{T}} \right\|_F^2 \leq \left\| \rho^t \mathbf{X}_t \right\|_F^2 ,$$

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.