Gossip Online Learning: Exchanging Local Models to Track Dynamics

January 9, 2019

Abstract

1 Introduction

For any online algorithm $A \in \mathcal{A}$, the previous dynamic regret $\widetilde{\mathcal{R}}_T^A$ is defined by

$$\widetilde{\mathcal{R}}_T^A = \sum_{i=1}^{n} \sum_{t=1}^{T} \left(g_{i,t}(\mathbf{x}_{i,t}) - g_{i,t}(\mathbf{x}_t^*) \right), \tag{1}$$

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

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

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 $t \in [T]$. $\mathbb{E}$ represents mathematical expectation. ∂ and ∇ represent sub-gradient and gradient operators, respectively. $\|\cdot\|$ represents the ℓ_2 norm in default.

3 Problem formulation

3.1 Setup

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

$$\mathcal{R}_T^A = \mathop{\mathbb{E}}_{\Xi_{n,T} \sim \mathcal{D}_{n,T}} \left(\sum_{i=1}^{n} \sum_{t=1}^{T} f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t}) - f_{i,t}(\mathbf{x}_t^*; \xi_{i,t}) \right), \tag{2}$$

where n is the number of nodes in the decentralized network. The local loss function $f_{i,t}(\mathbf{x}; \xi_{i,t})$ is defined by

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

with $0 < \beta < 1$, and $\xi_{i,t}$ is a random variable drawn from an unknown distribution $D_{i,t}$. Note that $g_{i,t}$ is an adversary loss function, which is yielded by the learning model. $h_t(\cdot; \xi_{i,t})$ is a known loss function, which depends on the random variable $\xi_{i,t}$. The expectation of $h_t(\cdot;\xi_{i,t})$ is a global model, and does not depend on the i-th node.

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

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

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

$$\sum_{t=1}^{T-1} \left\| \mathbf{x}_{t+1}^* - \mathbf{x}_t^* \right\| \le M. \tag{3}$$

When $M = 0$, all $\mathbf{x}_t^*$s are same, and it degenerates to the static online learning problem. When the dynamic environment changes significantly, M becomes large to model the dynamics. Besides, we denote

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

and

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

Recall that the previous definition of the dynamic regret is (1). Using (1), the classic online learning in a decentralized network only considers the loss function, i.e., $g_{i,t}$, incurred by the learning model on every node. Comparing with it, our definition of the dynamic regret, i.e., (2), still considers the loss function, i.e., H_t. It is incurred by a global model, which is used to let the decision variables, e.g., $\mathbf{x}_{i,t}$, have some good property in practical scenarios. We present some application scenarios to explain it in Section 3.2.

3.2 Application scenarios

Communication efficient online learning. Suppose we want to conduct online learning in a decentralized network. At every iteration, a node has to broadcast the local model to its neighbours, and the communication efficiency needs to be considered. In the case, $g_{i,t}(\mathbf{x})$ represents the loss incurred by the learning model, and $h_t(\mathbf{x};\xi_{i,t})$ represents the loss incurred by some a quantization method to guarantee the communication efficiency. A small β means a strong guarantee for the communication efficiency.

Suppose we want to conduct online classification by using logistic regression model. Given an instance $\mathbf{a}_{i,t} \in \mathbb{R}^d$ and its label $\mathbf{y}_{i,t} \in \{1,-1\}$. In the case, $g_{i,t}(\mathbf{x}) = \log\left(1+\exp(-\mathbf{y}_{i,t}\mathbf{a}_{i,t}^{\mathrm{T}}\mathbf{x})\right)$. We let $h_t(\mathbf{x};\xi_{i,t}) = \lambda_t \|\mathbf{Q}\mathbf{x}\|_1$[1]. Here, λ_t with $\lambda_t > 0$ is a given hyper-parameter. By using different λ_t over t, it is flexiable to adjust the communicaion efficiency timely. $\mathbf{Q} \in \mathbb{R}^{(d-1)\times d}$ is a special matrix:

$$\mathbf{Q} = \begin{bmatrix} 1 & -1 & & & \\ & 1 & -1 & & \\ & & \cdots & & \\ & & & 1 & -1 \end{bmatrix}.$$

Here, $h_t(\mathbf{x};\xi_{i,t})$ induces the difference between elements of $\mathbf{x}$ to be sparse. Thus, it is able to transmit $\mathbf{x}$ by using few different elements, and improve the communication efficiency. When λ_t is a constant, and does not change over t, $H_t(\mathbf{x})$ with $H_t(\mathbf{x}) = \lambda_t \|\mathbf{Q}\mathbf{x}\|_1$ plays a role of a regularizer.

Online learning with privacy protection. 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 encryption method to protect the local model

[1]In the case, the random variable $\xi_{i,t}$ is not necessary, which is a special case.

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., objective perturbation [??], to protect the privacy. A small β means a strong guarantee for the data privacy.

Similarly, suppose we want to conduct online classification by using logistic regression model. Given an instance $\mathbf{a}_{i,t}\in\mathbb{R}^d$ and its label $\mathbf{y}_{i,t}\in\{1,-1\}$. In the case, $g_{i,t}(\mathbf{x})=\log\left(1+\exp(-\mathbf{y}_{i,t}\mathbf{a}_{i,t}^{\mathrm{T}}\mathbf{x})\right)$. We use the objective perturbation strategy [??] to protect the privacy. Specifically, we let $h_t(\mathbf{x};\xi_{i,t})=\mathbf{x}^{\mathrm{T}}\xi_{i,t}$, where $\xi_{i,t}$ is random noise, whose density is

$$v(\mathbf{x})=\frac{1}{\lambda}\exp(-\delta_{i,t}\left\|\mathbf{x}\right\|).$$

Here, λ is a given hyper-parameter, $\delta_{i,t}$ is a known function of the constant $\epsilon_{i,t}$ for $\epsilon_{i,t}$-differential privacy [?]. In the case, $H_t(\mathbf{x})$ makes the decision variable own privacy-preserving property.

Online recommendation with unreliable features. Suppose we want to decide whether to recommend 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 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 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. A small β means significant attention on those unreliable features.

Suppose we still use logistic regression to decide whether to recommend music to Bob. Without loss of generality, features corresponding to those unreliable values are denoted by the beginning s features. Given a user's behavior record $\mathbf{a}_{i,t}$ and its label $\mathbf{y}_{i,t}\in\{1,-1\}$. In the case, $g_{i,t}(\mathbf{x})=\log\left(1+\exp\left(-\mathbf{y}_{i,t}\mathbf{a}_{i,t}^{\mathrm{T}}\hat{\mathbf{I}}\mathbf{x}\right)\right)$, where $\hat{\mathbf{I}}$ is yielded by letting the first s diagonal elements of an identity matrix be 0s. $\xi_{i,t}=\check{\mathbf{I}}\mathbf{a}_{i,t}\mathbf{y}_{i,t}^{\mathrm{T}}$, and $h_t(\mathbf{x};\xi_{i,t})=\log\left(1+\exp\left(-\xi_{i,t}^{\mathrm{T}}\mathbf{x}\right)\right)$, where $\check{\mathbf{I}}$ is yielded by letting the last $(d-s)$ diagonal elements of an identity matrix be 0s. Here, $\xi_{i,t}$ is drawn form an unknown distribution, that is, $\xi_{i,t}\sim D_{i,t}$, and $D_{i,t}$ usually changes unsignificant over t, or does not change over t. In the case, $H_t(\mathbf{x})$ allows the decision variable to represent different models to treat the unreliable and reliable features.

4 Algorithm

Algorithm 1 DOG: Decentralized Online Gradient method.

Require: The learning rate η, number of iterations T, and the confusion matrix $\mathbf{W}$. $\mathbf{x}_{i,1}=\mathbf{0}$ for any $i\in[n]$.

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

The decentralized online gradient method, namely DOG, is presented in Algorithm 1. At every iteration, every node needs to collect the decision variable, e.g., $\mathbf{x}_{i,t}$, from its neighbours, and then update its decision variable. Here, $\mathbf{W}\in\mathbb{R}^{n\times n}$ is the confusion matrix. It is a doublely stochastic matrix, which implies that every element of $\mathbf{W}$ is non-negative, $\mathbf{W}\mathbf{1}=\mathbf{1}$, and $\mathbf{1}^{\mathrm{T}}\mathbf{W}=\mathbf{1}^{\mathrm{T}}$. Denote $\bar{\mathbf{x}}_t=\frac{1}{n}\sum_{i=1}^n\mathbf{x}_{i,t}$. We can verify that $\bar{\mathbf{x}}_{t+1}=\bar{\mathbf{x}}_t-\frac{\eta}{n}\sum_{i=1}^n\partial f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t})$ (see Lemma 3).

5 Theoretical analysis

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*

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

  and

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

- *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, but may be non-smooth. Furthermore, we assume the function H_t has L-Lipschitz gradients. In brief, $g_{i,t}$ 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 $\bar{\mathbf{x}}_t = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_{i,t}$, and constants C_0 and C_1 by*

$$C_0 := \frac{1}{\sqrt{\beta^2 + \eta}} + 4;$$
$$C_1 := \frac{\beta}{2\eta} + L + \frac{\sqrt{\beta^2 + \eta}}{2\eta} + 2\eta L^2 + C_0(1-\beta)^2 L^2 \eta.$$

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

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\Xi_{n,T} \sim \mathcal{D}_{n,T}} \sum_{t=1}^T \sum_{i=1}^n f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t}) - f_{i,t}(\mathbf{x}_t^*; \xi_{i,t}) \\
\leq & \eta T \left(n\beta G + (1-\beta)\sigma^2\right) + n(1-\beta)C_0 \left(\mathop{\mathbb{E}}_{\Xi_{n,T} \sim \mathcal{D}_{n,T}} \sum_{t=1}^T \left(H_t(\bar{\mathbf{x}}_t) - H_t(\bar{\mathbf{x}}_{t+1})\right) \right) \\
&+ (1-\beta)\frac{nT\eta^2 G C_1}{(1-\rho)^2} + n(1-\beta)C_0 \left(4T\beta^2\eta G + \frac{TGL\eta^2}{2}\right) + \frac{n}{2\eta}\left(4\sqrt{R}M + R\right).
\end{aligned}$$

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

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

6 Empirical studies

6.1 Experimental settings

We simulate a decentralized network consisting of 5 nodes. Those nodes are connected by using a ring topology. Besides, we conduct online logistic regression by using time series data: *room-occupancy*[2], *online-retail*[3] in the decentralized network.

- *room-occupancy*.
- *online-retail*. It is an online retail dataset, which contains all transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. We use three features, that is, *whether a transaction is cancelled*, *quantity*, and *unit price*. We need to training a binary clssification model to make a decision whether a customer is coming from United Kingdom.

All values of a feature have been normalized to be zero mean and one variance. As we have shown in Section 3.2, we test the dynamic regret in those three application scenarios.

6.2 Communication efficient online logistic regression

6.3 Online logistic regression with privacy protection

6.4 Online logistic regression with unrealiable features

[2] https://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+

[3] https://archive.ics.uci.edu/ml/datasets/Online+Retail

Appendix

Proof to Theorem 1:

Proof.

$$
\begin{aligned}
& \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)}
\end{aligned}
$$

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

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

① holds due to $\langle \mathbf{a},\mathbf{b}\rangle \leq \frac{\eta}{2}\left\|\mathbf{a}\right\|^2 + \frac{1}{2\eta}\left\|\mathbf{b}\right\|^2$ 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

$$
\begin{aligned}
&J_1(t)\\
=&\frac{1}{n}\mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}}\sum_{i=1}^n\left\langle\nabla h_t(\mathbf{x}_{i,t};\xi_{i,t}),\mathbf{x}_{i,t}-\bar{\mathbf{x}}_t\right\rangle\\
=&\frac{1}{n}\mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}}\sum_{i=1}^n\left\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\right\rangle+\frac{1}{n}\mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}\sum_{i=1}^n\left\langle\nabla H_t(\bar{\mathbf{x}}_t),\mathbf{x}_{i,t}-\bar{\mathbf{x}}_t\right\rangle\\
=&\frac{1}{n}\mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}\sum_{i=1}^n\left\langle\nabla H_t(\mathbf{x}_{i,t})-\nabla H_t(\bar{\mathbf{x}}_t),\mathbf{x}_{i,t}-\bar{\mathbf{x}}_t\right\rangle+\frac{1}{n}\mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}\sum_{i=1}^n\left\langle\nabla H_t(\bar{\mathbf{x}}_t),\mathbf{x}_{i,t}-\bar{\mathbf{x}}_t\right\rangle\\
\overset{①}{\le}&\frac{L}{n}\mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}\sum_{i=1}^n\left\|\mathbf{x}_{i,t}-\bar{\mathbf{x}}_t\right\|^2+\frac{1}{n}\mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}\sum_{i=1}^n\left\langle\nabla H_t(\bar{\mathbf{x}}_t),\mathbf{x}_{i,t}-\bar{\mathbf{x}}_t\right\rangle\\
\overset{②}{\le}&\frac{L}{n}\mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}\sum_{i=1}^n\left\|\mathbf{x}_{i,t}-\bar{\mathbf{x}}_t\right\|^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}\left\|\nabla H_t(\bar{\mathbf{x}}_t)\right\|^2+\frac{\nu}{2\eta}\left\|\mathbf{x}_{i,t}-\bar{\mathbf{x}}_t\right\|^2\right)\\
\le&\frac{L}{n}\mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}\sum_{i=1}^n\left\|\mathbf{x}_{i,t}-\bar{\mathbf{x}}_t\right\|^2+\frac{\eta}{2\nu}\mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}\left\|\nabla H_t(\bar{\mathbf{x}}_t)\right\|^2+\frac{\nu}{2\eta n}\mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}\sum_{i=1}^n\left\|\mathbf{x}_{i,t}-\bar{\mathbf{x}}_t\right\|^2. \quad (4)
\end{aligned}
$$

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

For $J_2(t)$, we have

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

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

① holds due to

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

② holds due to H_t has L Lipschitz gradients.

Therefore, we obtain

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

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

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

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

Therefore, we have

$$\begin{aligned}
&\sum_{t=1}^{T}(I_1(t)+I_2(t))\\
\le&\eta T\left(\beta G+\frac{(1-\beta)\sigma^2}{n}\right)+(1-\beta)\left(\frac{\beta}{2n\eta}+\frac{L}{n}+\frac{\nu}{2n\eta}+\frac{2\eta L^2}{n}\right)\mathop{\mathbb{E}}_{\Xi_{n,T-1}\sim\mathcal{D}_{n,T-1}}\sum_{i=1}^{n}\sum_{t=1}^{T}\|\mathbf{x}_{i,t}-\bar{\mathbf{x}}_t\|^2\\
&+\frac{1}{2\eta}\mathop{\mathbb{E}}_{\Xi_{n,T}\sim\mathcal{D}_{n,T}}\sum_{t=1}^{T}\|\bar{\mathbf{x}}_t-\bar{\mathbf{x}}_{t+1}\|^2+\left(\frac{\eta}{2\nu}+2\eta\right)(1-\beta)\mathop{\mathbb{E}}_{\Xi_{n,T-1}\sim\mathcal{D}_{n,T-1}}\sum_{t=1}^{T}\|\nabla H_t(\bar{\mathbf{x}}_t)\|^2.
\end{aligned}$$

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

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

According to Lemma 3, we have

$$\bar{\mathbf{x}}_{t+1}=\bar{\mathbf{x}}_t-\eta\left(\frac{1}{n}\sum_{i=1}^{n}\partial f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t})\right). \tag{5}$$

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

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

It is trivial to verify that (5) 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

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

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

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

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

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

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

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

$$\begin{aligned}
I_3(t) =& \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \left\langle \frac{1}{n}\sum_{i=1}^{n}\partial f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t}), \bar{\mathbf{x}}_{t+1} - \mathbf{x}_t^*\right\rangle \\
\leq& \frac{1}{2\eta}\mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}}\left(\lim_{\tau\to 0^+}\tau\left\|\left(\mathbf{x}_t^* - \bar{\mathbf{x}}_{t+1}\right)\right\|^2 + 2\left\langle \mathbf{x}_t^* - \bar{\mathbf{x}}_{t+1}, \bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_t\right\rangle\right) \\
=& \frac{1}{\eta}\mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}}\left\langle \mathbf{x}_t^* - \bar{\mathbf{x}}_{t+1}, \bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_t\right\rangle \\
=& \frac{1}{2\eta}\mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}}\left(\left\|\mathbf{x}_t^* - \bar{\mathbf{x}}_t\right\|^2 - \left\|\mathbf{x}_t^* - \bar{\mathbf{x}}_{t+1}\right\|^2 - \left\|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\right\|^2\right).
\end{aligned} \tag{6}$$

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 (3).

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

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

$$
\begin{aligned}
&\overset{①}{\leq}\eta T\left(n\beta G+(1-\beta)\sigma^2\right)+n(1-\beta)\left(\frac{1}{\nu}+4\right)\left(\mathop{\mathbb{E}}_{\Xi_{n,T}\sim\mathcal{D}_{n,T}}\sum_{t=1}^{T}\left(H_t(\bar{\mathbf{x}}_t)-H_t(\bar{\mathbf{x}}_{t+1})\right)\right)\\
&\quad+(1-\beta)\left(\frac{\beta}{2\eta}+L+\frac{\nu}{2\eta}+2\eta L^2+\left(\frac{1}{\nu}+4\right)(1-\beta)^2L^2\eta\right)\mathop{\mathbb{E}}_{\Xi_{n,T}\sim\mathcal{D}_{n,T}}\sum_{t=1}^{T}\sum_{i=1}^{n}\left\|\bar{\mathbf{x}}_t-\mathbf{x}_{i,t}\right\|^2\\
&\quad+n(1-\beta)\left(\frac{1}{\nu}+4\right)\left(4T\beta^2\eta G+\frac{TGL\eta^2}{2}\right)+\frac{n}{2\eta}\left(4\sqrt{R}M+R\right)\\
&\overset{②}{\leq}\eta T\left(n\beta G+(1-\beta)\sigma^2\right)+n(1-\beta)\left(\frac{1}{\nu}+4\right)\left(\mathop{\mathbb{E}}_{\Xi_{n,T}\sim\mathcal{D}_{n,T}}\sum_{t=1}^{T}\left(H_t(\bar{\mathbf{x}}_t)-H_t(\bar{\mathbf{x}}_{t+1})\right)\right)\\
&\quad+(1-\beta)\left(\frac{\beta}{2\eta}+L+\frac{\nu}{2\eta}+2\eta L^2+\left(\frac{1}{\nu}+4\right)(1-\beta)^2L^2\eta\right)\frac{nT\eta^2G}{(1-\rho)^2}\\
&\quad+n(1-\beta)\left(\frac{1}{\nu}+4\right)\left(4T\beta^2\eta G+\frac{TGL\eta^2}{2}\right)+\frac{n}{2\eta}\left(4\sqrt{R}M+R\right).
\end{aligned}
$$

① holds due to Lemma 2. That is, we have

$$
\begin{aligned}
&\frac{\eta}{2}\mathop{\mathbb{E}}_{\Xi_{n,T-1}\sim\mathcal{D}_{n,T-1}}\sum_{t=1}^{T}\left\|\nabla H_t(\bar{\mathbf{x}}_t)\right\|^2 \qquad (7)\\
\leq&\mathop{\mathbb{E}}_{\Xi_{n,T}\sim\mathcal{D}_{n,T}}\sum_{t=1}^{T}\left(H_t(\bar{\mathbf{x}}_t)-H_t(\bar{\mathbf{x}}_{t+1})\right)+4T\beta^2\eta G+\frac{(1-\beta)^2L^2\eta}{n}\mathop{\mathbb{E}}_{\Xi_{n,T-1}\sim\mathcal{D}_{n,T-1}}\sum_{t=1}^{T}\sum_{i=1}^{n}\left\|\bar{\mathbf{x}}_t-\mathbf{x}_{i,t}\right\|^2+\frac{TGL\eta^2}{2}.
\end{aligned}
$$

② holds due to Lemma 4

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

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

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

It completes the proof. □

Lemma 1. *Using Assumption 1, we have*

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

Proof.

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

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

It completes the proof. □

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

$$
\begin{aligned}
&\frac{\eta}{2} \mathop{\mathbb{E}}_{\Xi_{n,T-1}\sim\mathcal{D}_{n,T-1}} \sum_{t=1}^{T} \left\|\nabla H_t(\bar{\mathbf{x}}_t)\right\|^2 \qquad (8)\\
\le& \mathop{\mathbb{E}}_{\Xi_{n,T}\sim\mathcal{D}_{n,T}} \sum_{t=1}^{T} \left(H_t(\bar{\mathbf{x}}_t) - H_t(\bar{\mathbf{x}}_{t+1})\right) + 4T\beta^2\eta G + \frac{(1-\beta)^2 L^2 \eta}{n} \mathop{\mathbb{E}}_{\Xi_{n,T-1}\sim\mathcal{D}_{n,T-1}} \sum_{t=1}^{T}\sum_{i=1}^{n} \left\|\bar{\mathbf{x}}_t - \mathbf{x}_{i,t}\right\|^2 + \frac{TGL\eta^2}{2}.
\end{aligned}
$$

Proof.

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} H_t(\bar{\mathbf{x}}_{t+1}) \\
\le& \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}} H_t(\bar{\mathbf{x}}_t) + \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \left\langle \nabla H_t(\bar{\mathbf{x}}_t), \bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_t \right\rangle + \frac{L}{2} \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \left\|\bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_t\right\|^2 \\
=& \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}} H_t(\bar{\mathbf{x}}_t) + \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \left\langle \nabla H_t(\bar{\mathbf{x}}_t), -\frac{\eta}{n}\sum_{i=1}^{n}\partial f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t}) \right\rangle + \frac{L}{2} \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \left\|\frac{\eta}{n}\sum_{i=1}^{n}\partial f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t})\right\|^2 \\
=& \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}} H_t(\bar{\mathbf{x}}_t) + \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}} \left\langle \nabla H_t(\bar{\mathbf{x}}_t), -\frac{\eta}{n}\sum_{i=1}^{n}\partial f_{i,t}(\mathbf{x}_{i,t}) \right\rangle + \frac{L}{2} \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \left\|\frac{\eta}{n}\sum_{i=1}^{n}\partial f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t})\right\|^2.
\end{aligned}
\tag{9}
$$

Besides, we have

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

$$
\begin{aligned}
&\leq \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}} \frac{\eta}{2}\left(2\beta^2\left\|\nabla H_t(\bar{\mathbf{x}}_t)-\frac{1}{n}\sum_{i=1}^n \partial g_{i,t}(\mathbf{x}_{i,t})\right\|^2+\frac{2(1-\beta)^2L^2}{n}\sum_{i=1}^n\left\|\bar{\mathbf{x}}_t-\mathbf{x}_{i,t}\right\|^2\right)-\mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}\frac{\eta}{2}\left\|\nabla H_t(\bar{\mathbf{x}}_t)\right\|^2\\
&\leq \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}} \frac{\eta}{2}\left(4\beta^2\left\|\nabla H_t(\bar{\mathbf{x}}_t)\right\|^2+4\beta^2\left\|\frac{1}{n}\sum_{i=1}^n \partial g_{i,t}(\mathbf{x}_{i,t})\right\|^2+\frac{2(1-\beta)^2L^2}{n}\sum_{i=1}^n\left\|\bar{\mathbf{x}}_t-\mathbf{x}_{i,t}\right\|^2\right)\\
&\quad-\mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}\frac{\eta}{2}\left\|\nabla H_t(\bar{\mathbf{x}}_t)\right\|^2\\
&\overset{①}{\leq} \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}} \frac{\eta}{2}\left(8\beta^2G+\frac{2(1-\beta)^2L^2}{n}\sum_{i=1}^n\left\|\bar{\mathbf{x}}_t-\mathbf{x}_{i,t}\right\|^2\right)-\mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}}\frac{\eta}{2}\left\|\nabla H_t(\bar{\mathbf{x}}_t)\right\|^2. \qquad (10)
\end{aligned}
$$

① holds due to

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

and

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

According to Lemma 1, we have

$$
\mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}}\left\|\partial f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t})\right\|^2 \leq G. \qquad (11)
$$

Substituting (10) and (11) into (9), and telescoping $t\in[T]$, we obtain

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\Xi_{n,T}\sim\mathcal{D}_{n,T}}\sum_{t=1}^T H_t(\bar{\mathbf{x}}_{t+1})\\
&\leq \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}} H_t(\bar{\mathbf{x}}_t)+\mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}\left\langle \nabla H_t(\bar{\mathbf{x}}_t), -\frac{\eta}{n}\sum_{i=1}^n \partial f_{i,t}(\mathbf{x}_{i,t})\right\rangle+\frac{L}{2}\mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}}\left\|\frac{\eta}{n}\sum_{i=1}^n \partial f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t})\right\|^2\\
&\leq \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}} H_t(\bar{\mathbf{x}}_t)+\left(\mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}\frac{\eta}{2}\left(8\beta^2G+\frac{2(1-\beta)^2L^2}{n}\sum_{i=1}^n\left\|\bar{\mathbf{x}}_t-\mathbf{x}_{i,t}\right\|^2\right)-\mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}\frac{\eta}{2}\left\|\nabla H_t(\bar{\mathbf{x}}_t)\right\|^2\right)+\frac{GL\eta^2}{2}\\
&= \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}} H_t(\bar{\mathbf{x}}_t)+\left(4\eta\beta^2G+\frac{(1-\beta)^2L^2\eta}{n}\mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}\sum_{i=1}^n\left\|\bar{\mathbf{x}}_t-\mathbf{x}_{i,t}\right\|^2-\mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}\frac{\eta}{2}\left\|\nabla H_t(\bar{\mathbf{x}}_t)\right\|^2\right)+\frac{GL\eta^2}{2}.
\end{aligned}
$$

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

$$
\begin{aligned}
&\frac{\eta}{2}\mathop{\mathbb{E}}_{\Xi_{n,T-1}\sim\mathcal{D}_{n,T-1}}\sum_{t=1}^T\left\|\nabla H_t(\bar{\mathbf{x}}_t)\right\|^2 \qquad (12)\\
&\leq \mathop{\mathbb{E}}_{\Xi_{n,T}\sim\mathcal{D}_{n,T}}\sum_{t=1}^T\left(H_t(\bar{\mathbf{x}}_t)-H_t(\bar{\mathbf{x}}_{t+1})\right)+4T\beta^2\eta G+\frac{(1-\beta)^2L^2\eta}{n}\mathop{\mathbb{E}}_{\Xi_{n,T-1}\sim\mathcal{D}_{n,T-1}}\sum_{t=1}^T\sum_{i=1}^n\left\|\bar{\mathbf{x}}_t-\mathbf{x}_{i,t}\right\|^2+\frac{TGL\eta^2}{2}.
\end{aligned}
$$

It completes the proof. □

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

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

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 doublely 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*

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

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

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

Denote

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

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

$$\bar{\mathbf{x}}_{t+1} = \bar{\mathbf{x}}_t - \eta\left(\frac{1}{n}\sum_{i=1}^{n}\partial f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t})\right) = -\sum_{s=1}^{t}\eta\bar{\mathbf{G}}_s. \tag{13}$$

Therefore,

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

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

Thus, we have

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

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

□

Lemma 5 (Appeared in Lemma 5 in [**?**]). *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*

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

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

Lemma 6 (Appeared in Lemma 6 in [**?**]). *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*

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