Gossip Online Learning: Exchanging Local Models to Tracking Dynamics

January 2, 2019

Abstract

ddd

1 Problem setup

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

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

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

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

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

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

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

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

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

The budget of the dynamics is defined as

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

Algorithm 1 DOG: Decentralized Online Gradient.

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

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

2 Algorithm

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

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

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

3 Theoretical analysis

3.1 Assumptions

Assumption 1. *We make the following assumptions.*

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

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

  and

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

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

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

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

Recall that $H_t(\cdot) = \mathbb{E}_{\xi_{i,t} \sim D_{i,t}}\, h_t(\cdot; \xi_{i,t})$. Assumption 2 implies that the cumulative difference between two successive distributions, e.g., $D_{i,t}$ and $D_{i,t+1}$, cannot be arbitrary.

Theorem 1. *Denote*

$$
\begin{aligned}
C_0 :=& n\left(\frac{1}{\beta+\eta}+4\right);\\
C_1 :=& L+\frac{\eta L^2}{\beta+\eta}+\frac{2\beta+\eta}{2\eta}+6\eta L^2;\\
C_2 :=& \frac{T\beta(\eta+8nL\eta^2+16n\eta\beta^2)+2nTL\eta^2}{\beta+\eta}.
\end{aligned}
$$

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

$$
\begin{aligned}
\mathcal{R}_T^{DOG} =& \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \sum_{t=1}^{T}\sum_{i=1}^{n} f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t}) - f_t(\mathbf{x}_t^*;\xi_{i,t})\\
\leq& 31nT\beta G\eta + 4(1-\beta)nT\eta\beta\sigma^2 + \frac{\eta^2 C_1}{(1-\rho)^2}\left(6nTG+4(1-\beta)nT\sigma^2\right)\\
&+(1-\beta)C_0 \mathop{\mathbb{E}}_{\Xi_{n,T}\sim\mathcal{D}_{n,T}} \sum_{t=1}^{T}\left(H_t(\bar{\mathbf{x}}_t)-H_t(\bar{\mathbf{x}}_{t+1})\right) + 3TLG\eta^2 C_0\\
&+(1-\beta)C_2\sigma^2 + \frac{n}{2\eta}\left(4\sqrt{R}M+R\right).
\end{aligned}
$$

Corollary 1. *Recall that $C_0 = n\left(\frac{1}{\beta+\eta}+4\right)$. Using Assumption 1, and choosing*

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

in Algorithm 1, we have

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

Appendix

Proof to Theorem 1:

Proof.

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

$$
\begin{aligned}
&+\frac{1}{n}\mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}}\sum_{i=1}^{n}(1-\beta)\left(\langle\nabla h_t(\mathbf{x}_{i,t};\xi_{i,t}),\bar{\mathbf{x}}_{t+1}-\mathbf{x}_t^*\rangle\right)\\
=&\underbrace{\mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}}\frac{1}{n}\sum_{i=1}^{n}\beta\left(\langle\nabla g_{i,t}(\mathbf{x}_{i,t}),\mathbf{x}_{i,t}-\bar{\mathbf{x}}_t\rangle+\langle\nabla g_{i,t}(\mathbf{x}_{i,t}),\bar{\mathbf{x}}_t-\bar{\mathbf{x}}_{t+1}\rangle\right)}_{I_1(t)}\\
&+\underbrace{\mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}}\frac{1}{n}\sum_{i=1}^{n}(1-\beta)\left(\langle\nabla h_t(\mathbf{x}_{i,t};\xi_{i,t}),\mathbf{x}_{i,t}-\bar{\mathbf{x}}_t\rangle+\langle\nabla h_t(\mathbf{x}_{i,t};\xi_{i,t}),\bar{\mathbf{x}}_t-\bar{\mathbf{x}}_{t+1}\rangle\right)}_{I_2(t)}\\
&+\underbrace{\mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}}\left\langle\frac{1}{n}\sum_{i=1}^{n}\nabla f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t}),\bar{\mathbf{x}}_{t+1}-\mathbf{x}_t^*\right\rangle}_{I_3(t)}
\end{aligned}
$$

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

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

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

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

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

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

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

$$\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 \\
\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} \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}} \|\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} \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}} \|\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} \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}} \|\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} \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}} \|\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{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 .
\end{aligned} \tag{3}$$

① holds due to

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

② holds due to H_t has L Lipschitz gradients.

Therefore, we obtain

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

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

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

Therefore, we have

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

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

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

According to Lemma 3, we have

$$\bar{\mathbf{x}}_{t+1} = \bar{\mathbf{x}}_t - \eta \left(\frac{1}{n} \sum_{i=1}^{n} \nabla f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t}) \right). \tag{4}$$

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

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

It is trivial to verify that (4) satisfies the first-order optimality condition of the optimization problem: $\min_{\mathbf{z}\in\mathbb{R}^d} \phi(\mathbf{z})$, that is,

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

We thus have

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

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

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

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

$$\begin{aligned} I_3(t) =& \mathop{\mathbb{E}}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \left\langle \frac{1}{n} \sum_{i=1}^{n} \nabla f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t}), \bar{\mathbf{x}}_{t+1} - \mathbf{x}_t^* \right\rangle \\ \le& \frac{1}{2\eta} \mathop{\mathbb{E}}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \left(\lim_{\tau \to 0^+} \tau \left\| \left(\mathbf{x}_t^* - \bar{\mathbf{x}}_{t+1} \right) \right\|^2 + 2 \left\langle \mathbf{x}_t^* - \bar{\mathbf{x}}_{t+1}, \bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_t \right\rangle \right) \\ =& \frac{1}{\eta} \mathop{\mathbb{E}}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \left\langle \mathbf{x}_t^* - \bar{\mathbf{x}}_{t+1}, \bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_t \right\rangle \\ =& \frac{1}{2\eta} \mathop{\mathbb{E}}_{\Xi_{n,t} \sim \mathcal{D}_{n,t}} \left(\left\| \mathbf{x}_t^* - \bar{\mathbf{x}}_t \right\|^2 - \left\| \mathbf{x}_t^* - \bar{\mathbf{x}}_{t+1} \right\|^2 - \left\| \bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1} \right\|^2 \right). \end{aligned} \tag{5}$$

Besides, we have

$$\left\| \mathbf{x}_{t+1}^* - \bar{\mathbf{x}}_{t+1} \right\|^2 - \left\| \mathbf{x}_t^* - \bar{\mathbf{x}}_{t+1} \right\|^2$$

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

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

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

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

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

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

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\Xi_{n,T}\sim\mathcal{D}_{n,T}} \sum_{t=1}^T \sum_{i=1}^n f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t}) - f_t(\mathbf{x}_t^*;\xi_{i,t}) \\
\le& n\sum_{t=1}^T \left(I_1(t) + I_2(t) + I_3(t)\right) \\
\le& \eta T\left(n\beta G + (1-\beta)\sigma^2\right) + \left(\frac{\beta}{2\eta} + L + \frac{\nu}{2\eta} + 2\eta L^2\right) \mathop{\mathbb{E}}_{\Xi_{n,T-1}\sim\mathcal{D}_{n,T-1}} \sum_{i=1}^n \sum_{t=1}^T \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 \\
&+ n\left(\frac{\eta}{2\nu} + 2\eta\right)(1-\beta) \mathop{\mathbb{E}}_{\Xi_{n,T-1}\sim\mathcal{D}_{n,T-1}} \sum_{t=1}^T \|\nabla H_t(\bar{\mathbf{x}}_t)\|^2 + \frac{n}{2\eta}\left(4\sqrt{R}M + R\right) \\
\overset{①}{\le}& \eta T\left(n\beta G + (1-\beta)\sigma^2\right) + \left(\frac{\beta}{2\eta} + L + \frac{\nu}{2\eta} + 2\eta L^2\right) \frac{4nT\eta^2 G}{(1-\rho)^2} + \frac{n}{2\eta}\left(4\sqrt{R}M + R\right) \\
&+ n\left(\frac{2}{\nu} + 8\right)\left(\mathop{\mathbb{E}}_{\Xi_{n,T}\sim\mathcal{D}_{n,T}} \sum_{t=1}^T \left(H_t(\bar{\mathbf{x}}_t) - H_t(\bar{\mathbf{x}}_{t+1})\right) + \frac{T\eta\beta^2 G}{1-\beta} + \frac{2GT\eta^3(1-\beta)L^2}{(1-\rho)^2} + 2LT\eta^2 G \right) \\
=& \eta T\left(n\beta G + (1-\beta)\sigma^2\right) + \frac{n}{2\eta}\left(4\sqrt{R}M + R\right) + n\left(\frac{2}{\nu} + 8\right) \mathop{\mathbb{E}}_{\Xi_{n,T}\sim\mathcal{D}_{n,T}} \sum_{t=1}^T \left(H_t(\bar{\mathbf{x}}_t) - H_t(\bar{\mathbf{x}}_{t+1})\right) \\
&+ \left(\frac{2nT\eta\beta}{(1-\rho)^2} + \frac{4nT\eta^2 L}{(1-\rho)^2} + \frac{2nT\eta\nu}{(1-\rho)^2} + \frac{8nT\eta^3 L^2}{(1-\rho)^2} + \left(\frac{2}{\nu} + 8\right)\left(\frac{nT\eta\beta^2}{1-\beta} + \frac{2nT\eta^3(1-\beta)L^2}{(1-\rho)^2} + 2nLT\eta^2 \right) \right) G.
\end{aligned}
$$

① holds due to Lemma 2 and Lemma 4. That is, we have

$$
\begin{aligned}
&\frac{\eta(1-\beta)}{4} \sum_{t=1}^T \|\nabla H_t(\bar{\mathbf{x}}_t)\|^2 \\
\le& \mathop{\mathbb{E}}_{\Xi_{n,T-1}\sim\mathcal{D}_{n,T-1}} \sum_{t=1}^T \left(H_t(\bar{\mathbf{x}}_t) - H_t(\bar{\mathbf{x}}_{t+1})\right) + \frac{T\eta\beta^2 G}{1-\beta} + \frac{2GT\eta^3(1-\beta)L^2}{(1-\rho)^2} + 2LT\eta^2 G.
\end{aligned} \tag{6}
$$

and

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

It completes the proof.

$\square$

Lemma 1. *Using Assumption 1, we have*

$$\mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \left\|\nabla f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t})\right\|^2 \le 4G.$$

Proof.

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

$\square$

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

$$\begin{aligned}
&\frac{\eta(1-\beta)}{4}\sum_{t=1}^{T}\left\|\nabla H_t(\bar{\mathbf{x}}_t)\right\|^2 \\
\le&\mathop{\mathbb{E}}_{\Xi_{n,T-1}\sim\mathcal{D}_{n,T-1}}\sum_{t=1}^{T}\left(H_t(\bar{\mathbf{x}}_t)-H_t(\bar{\mathbf{x}}_{t+1})\right)+\frac{T\eta\beta^2G}{1-\beta}+\frac{2GT\eta^3(1-\beta)L^2}{(1-\rho)^2}+2LT\eta^2G.
\end{aligned}\tag{7}$$

Proof.

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

Besides, we have

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

$$
\begin{aligned}
=&\eta \underset{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}{\mathbb{E}} \left\langle \nabla H_t(\bar{\mathbf{x}}_t), -\frac{\beta}{n}\sum_{i=1}^{n}\nabla g_{i,t}(\mathbf{x}_{i,t})\right\rangle \\
\le&\frac{\eta(1-\beta)}{4}\underset{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}{\mathbb{E}}\left\|\nabla H_t(\bar{\mathbf{x}}_t)\right\|^2 + \frac{\eta\beta^2}{1-\beta}\underset{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}{\mathbb{E}}\left\|\frac{1}{n}\sum_{i=1}^{n}\nabla g_{i,t}(\mathbf{x}_{i,t})\right\|^2 \\
\le&\frac{\eta(1-\beta)}{4}\underset{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}{\mathbb{E}}\left\|\nabla H_t(\bar{\mathbf{x}}_t)\right\|^2 + \frac{\eta\beta^2}{1-\beta}\frac{1}{n}\underset{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}{\mathbb{E}}\sum_{i=1}^{n}\left\|\nabla g_{i,t}(\mathbf{x}_{i,t})\right\|^2 \\
\le&\frac{\eta(1-\beta)}{4}\underset{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}}{\mathbb{E}}\left\|\nabla H_t(\bar{\mathbf{x}}_t)\right\|^2 + \frac{\eta\beta^2 G}{1-\beta},
\end{aligned}
$$

and

$$
\begin{aligned}
&\langle \nabla H_t(\bar{\mathbf{x}}_t), -\nabla H_t(\mathbf{x}_{i,t})\rangle \\
=&\frac{1}{2}\left(\left\|\nabla H_t(\bar{\mathbf{x}}_t)-\nabla H_t(\mathbf{x}_{i,t})\right\|^2 - \left\|\nabla H_t(\bar{\mathbf{x}}_t)\right\|^2 - \left\|\nabla H_t(\mathbf{x}_{i,t})\right\|^2\right) \\
\le&\frac{1}{2}\left(L^2\left\|\bar{\mathbf{x}}_t-\mathbf{x}_{i,t}\right\|^2 - \left\|\nabla H_t(\bar{\mathbf{x}}_t)\right\|^2 - \left\|\nabla H_t(\mathbf{x}_{i,t})\right\|^2\right),
\end{aligned}
$$

and

$$
\underset{\Xi_{n,t}\sim\mathcal{D}_{n,t}}{\mathbb{E}}\left\|\nabla f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t})\right\|^2 \le 4G.
$$

Thus, telescoping $t\in[T]$, we obtain

$$
\begin{aligned}
&\underset{\Xi_{n,T}\sim\mathcal{D}_{n,T}}{\mathbb{E}}\sum_{t=1}^{T}H_t(\bar{\mathbf{x}}_{t+1}) \qquad (8)\\
\le&\underset{\Xi_{n,T-1}\sim\mathcal{D}_{n,T-1}}{\mathbb{E}}\sum_{t=1}^{T}H_t(\bar{\mathbf{x}}_t) + \frac{\eta(1-\beta)}{4}\underset{\Xi_{n,T-1}\sim\mathcal{D}_{n,T-1}}{\mathbb{E}}\sum_{t=1}^{T}\left\|\nabla H_t(\bar{\mathbf{x}}_t)\right\|^2 + \frac{T\eta\beta^2 G}{1-\beta} \\
&+\frac{\eta(1-\beta)L^2}{2n}\sum_{t=1}^{T}\sum_{i=1}^{n}\left\|\bar{\mathbf{x}}_t-\mathbf{x}_{i,t}\right\|^2 - \frac{\eta(1-\beta)}{2}\sum_{t=1}^{T}\left\|\nabla H_t(\bar{\mathbf{x}}_t)\right\|^2 \\
&-\frac{\eta(1-\beta)}{2n}\sum_{i=1}^{n}\sum_{t=1}^{T}\left\|\nabla H_t(\mathbf{x}_{i,t})\right\|^2 + 2LT\eta^2 G \\
\le&\underset{\Xi_{n,T-1}\sim\mathcal{D}_{n,T-1}}{\mathbb{E}}\sum_{t=1}^{T}H_t(\bar{\mathbf{x}}_t) + \frac{T\eta\beta^2 G}{1-\beta} + \frac{2GT\eta^3(1-\beta)L^2}{(1-\rho)^2} - \frac{\eta(1-\beta)}{4}\sum_{t=1}^{T}\left\|\nabla H_t(\bar{\mathbf{x}}_t)\right\|^2 \\
&-\frac{\eta(1-\beta)}{2n}\sum_{i=1}^{n}\sum_{t=1}^{T}\left\|\nabla H_t(\mathbf{x}_{i,t})\right\|^2 + 2LT\eta^2 G.
\end{aligned}
$$

The last inequality holds due to Lemma 4, that is,

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

Equivalently, we have

$$
\frac{\eta(1-\beta)}{4}\sum_{t=1}^{T}\left\|\nabla H_t(\bar{\mathbf{x}}_t)\right\|^2 \qquad (9)
$$

$$
\begin{aligned}
\le & \frac{\eta(1-\beta)}{4}\sum_{t=1}^{T}\|\nabla H_t(\bar{\mathbf{x}}_t)\|^2 + \frac{\eta(1-\beta)}{2n}\sum_{i=1}^{n}\sum_{t=1}^{T}\|\nabla H_t(\mathbf{x}_{i,t})\|^2 \\
\le & \mathop{\mathbb{E}}_{\Xi_{n,T-1}\sim\mathcal{D}_{n,T-1}}\sum_{t=1}^{T}\left(H_t(\bar{\mathbf{x}}_t) - H_t(\bar{\mathbf{x}}_{t+1})\right) + \frac{T\eta\beta^2 G}{1-\beta} + \frac{2GT\eta^3(1-\beta)L^2}{(1-\rho)^2} + 2LT\eta^2 G.
\end{aligned}
$$

It completes the proof.

□

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

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

Proof. Denote

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

Recall that

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

Equivalently, we re-formulate the update rule as

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

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

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

Thus, we have

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

□

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

$$
\mathop{\mathbb{E}}_{\Xi_{n,T}\sim\mathcal{D}_{n,T}}\sum_{i=1}^{n}\sum_{t=1}^{T}\|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 \le \frac{4nT\eta^2 G}{(1-\rho)^2}.
$$

Proof. Recall that

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

and according to Lemma 3, we have

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

Denote

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

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

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

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

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

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

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

$$
\begin{aligned}
&\overset{①}{\leq} \frac{\eta^2}{(1-\rho)^2} \underset{\Xi_{n,T}\sim\mathcal{D}_{n,T}}{\mathbb{E}} \left(\sum_{t=1}^{T} \|\mathbf{G}_t\|_F^2\right) \\
&= \frac{\eta^2}{(1-\rho)^2} \left(\underset{\Xi_{n,T}\sim\mathcal{D}_{n,T}}{\mathbb{E}} \sum_{t=1}^{T}\sum_{i=1}^{n} \left\|\nabla f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t})\right\|^2\right) \\
&\overset{②}{=} \frac{4nT\eta^2 G}{(1-\rho)^2}.
\end{aligned}
$$

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 $\mathbf{W} = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^{\mathrm{T}} = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^{\mathrm{T}}$, 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. $\boldsymbol{\Lambda}$ 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 [?]). *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

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