Decentralized Online Optimization

December 26, 2018

Abstract

ddd

1 Notations and assumptions

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 $\mathbf{x}$,

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

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. $h_{i,t}(\cdot,\xi_{i,t})$ is a given loss function depending on the random variable $\xi_{i,t}$. Besides, we denote

$$H_{i,t}(\cdot) = \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} h_{i,t}(\cdot;\xi_{i,t}).$$

The budget of the dynamics is defined as

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

Assumption 1. *We make the following assumptions.*

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

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

and

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

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

2 Algorithm

Algorithm 1 DOG: Decentralized Online Gradient.

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

```
1: for t = 1, 2, ..., T do
     For the i-th node with i ∈ [n]:
2:       Predict x_{i,t}.
3:       Observe the loss function f_{i,t},
         and suffer loss f_{i,t}(x_{i,t}; ξ_{i,t}).
     Update:
4:       Query the gradient ∇f_{i,t}(x_{i,t}; ξ_{i,t}).
5:       x_{i,t+1} = Σ_{j=1}^{n} W_{i,j} x_{j,t} − η∇f_{i,t}(x_{i,t}; ξ_{i,t}).
```

Theorem 1. *For any* $0 \le \beta \le 1$ *and* $\eta > 0$*, we denote*

$$\begin{aligned}C_1 =& 11\beta + \frac{6\eta}{(1-\rho)^2}\left(L + \frac{\eta L^2}{\beta} + 6\eta L^2 + \frac{\beta}{\eta}\right) + 3L\eta\left(\frac{1}{\beta}+4\right),\\ C_2 =& \left(L + \frac{\eta L^2}{\beta} + 6\eta L^2 + \frac{\beta}{\eta}\right)\frac{4\eta^2}{(1-\rho)^2} + \frac{\nu(\eta + 8nL\eta^2) + 2nL\eta^2}{n\beta}.\end{aligned}$$

Use Assumption 1, and choose $\eta > 0$ *in Algorithm 1. We have*

$$\mathcal{R}_T^{DOG} \le \eta n T G C_1 + \left(\frac{1}{\beta}+4\right) n \left(H_t(\bar{\mathbf{x}}_1) - H_t(\bar{\mathbf{x}}_{t+1})\right) + C_2 \sum_{t=1}^{T}\sum_{i=1}^{n} \sigma_{i,t}^2 + \frac{n}{2\eta}\left(4\sqrt{R}M + R\right).$$

Corollary 1. *Using Assumption 1, and choosing*

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

in Algorithm 1, we have

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

Appendix

Proof to Theorem 1:

Proof.

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

$$
\begin{aligned}
&\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}}\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) \\
&\quad + \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) \\
&\quad + \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}), \bar{\mathbf{x}}_{t+1}-\mathbf{x}_t^*\right\rangle\right) \\
&= \underbrace{\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\right)}_{I_1(t)} \\
&\quad + \underbrace{\mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \frac{1}{n}\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)}_{I_2(t)} \\
&\quad + \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 $I_1(t)$.

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

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

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

For $J_1(t)$ and $\nu > 0$, we have

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

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

① holds due to H_t has L-Lipischitz gradients.

According to Lemma 2, we have

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

Substituting (2) into (1), we obtain

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

For $J_2(t)$, we have

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

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

The last inequality holds due to H_t has L Lipschitz gradients.

Substituting (2) into (3), and we have

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

Therefore, we obtain

$$
\begin{aligned}
& I_2(t) \\
= & (1-\beta)(J_1(t) + J_2(t)) \\
= & (1-\beta)\left(\frac{L}{n} + \frac{\eta L^2}{n\nu} + \frac{\nu}{2n\eta} + \frac{6\eta L^2}{n}\right) \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{n,t-1}} \sum_{i=1}^n \left\| \mathbf{x}_{i,t} - \bar{\mathbf{x}}_t \right\|^2 + (1-\beta)\left(\frac{1}{\nu} + 4\right) \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \left(H_t(\bar{\mathbf{x}}_t) - H_t(\bar{\mathbf{x}}_{t+1}) \right) \\
& + (1-\beta)\left(2G\eta\beta^2\left(\frac{1}{\nu} + 4\right) + \frac{1}{2\eta} \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \left\| \bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1} \right\|^2\right) \\
& + \frac{(1-\beta)3LG\eta^2}{\nu}\left(\frac{1}{\nu} + 4\right) + (1-\beta)\frac{\nu(\eta + 8nL\eta^2) + 2nL\eta^2}{n^2\nu}\sum_{i=1}^n \sigma_{i,t}^2
\end{aligned}
$$

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

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

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

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

Therefore, we have

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

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

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

By taking average over $i \in [n]$ on both sides, 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 $h(\mathbf{z})$ as

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

Note that (4) is equivalent to

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

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

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

Dividing τ on both sides, and letting τ be close to 0, 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

$$\begin{aligned}&\left\|\mathbf{x}_{t+1}^* - \bar{\mathbf{x}}_{t+1}\right\|^2 - \left\|\mathbf{x}_t^* - \bar{\mathbf{x}}_{t+1}\right\|^2\\ =& \left\|\mathbf{x}_{t+1}^*\right\|^2 - \left\|\mathbf{x}_t^*\right\|^2 - 2\left\langle \bar{\mathbf{x}}_{t+1}, -\mathbf{x}_t^* + \mathbf{x}_{t+1}^*\right\rangle\\ =& \left(\left\|\mathbf{x}_{t+1}^*\right\| - \left\|\mathbf{x}_t^*\right\|\right)\left(\left\|\mathbf{x}_{t+1}^*\right\| + \left\|\mathbf{x}_t^*\right\|\right) - 2\left\langle \bar{\mathbf{x}}_{t+1}, -\mathbf{x}_t^* + \mathbf{x}_{t+1}^*\right\rangle\\ \le& \left\|\mathbf{x}_{t+1}^* - \mathbf{x}_t^*\right\|\left(\left\|\mathbf{x}_{t+1}^*\right\| + \left\|\mathbf{x}_t^*\right\|\right) - 2\left\|\bar{\mathbf{x}}_{t+1}\right\|\left\|\mathbf{x}_{t+1}^* - \mathbf{x}_t^*\right\|\\ \le& 4\sqrt{R}\left\|\mathbf{x}_{t+1}^* - \mathbf{x}_t^*\right\|. \quad \text{(due to } \left\|\mathbf{x}-\mathbf{y}\right\|^2 \le R, \forall \mathbf{x},\mathbf{y}\in\mathcal{X}\text{)}\end{aligned}$$

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

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

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

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

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\Xi_{n,T}\sim\mathcal{D}_{n,T}}\sum_{t=1}^{T}\sum_{i=1}^{n} f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t}) - f_t(\mathbf{x}_t^*;\xi_{i,t})\\
\leq & n\sum_{t=1}^{T}\left(I_1(t)+I_2(t)+I_3(t)\right)\\
\leq & \left(1+2\beta\left(\frac{1}{\nu}+4\right)\right)nT\beta G\eta + \left(L+\frac{\eta L^2}{\nu}+\frac{\nu}{2\eta}+6\eta L^2+\frac{\beta}{2\eta}\right)\mathop{\mathbb{E}}_{\Xi_{n,T}\sim\mathcal{D}_{n,T}}\sum_{t=1}^{T}\sum_{i=1}^{n}\left\|\mathbf{x}_{i,t}-\bar{\mathbf{x}}_t\right\|^2\\
&+\left(\frac{1}{\nu}+4\right)n\left(H_t(\bar{\mathbf{x}}_1)-H_t(\bar{\mathbf{x}}_{t+1})\right)\\
&+3nTLG\eta^2\left(\frac{1}{\nu}+4\right)+\frac{\nu(\eta+8nL\eta^2)+2nL\eta^2}{n\nu}\sum_{t=1}^{T}\sum_{i=1}^{n}\sigma_{i,t}^2+\frac{n}{2\eta}\left(4\sqrt{R}M+R\right)\\
\overset{①}{\leq} & \left(1+2\beta\left(\frac{1}{\nu}+4\right)\right)nT\beta G\eta + \left(L+\frac{\eta L^2}{\nu}+\frac{\nu}{2\eta}+6\eta L^2+\frac{\beta}{2\eta}\right)\frac{\eta^2}{(1-\rho)^2}\left(6nTG+4\sum_{t=1}^{T}\sum_{i=1}^{n}\sigma_{i,t}^2\right)\\
&+\left(\frac{1}{\nu}+4\right)n\left(H_t(\bar{\mathbf{x}}_1)-H_t(\bar{\mathbf{x}}_{t+1})\right)\\
&+3nTLG\eta^2\left(\frac{1}{\nu}+4\right)+\frac{\nu(\eta+8nL\eta^2)+2nL\eta^2}{n\nu}\sum_{t=1}^{T}\sum_{i=1}^{n}\sigma_{i,t}^2+\frac{n}{2\eta}\left(4\sqrt{R}M+R\right).
\end{aligned}$$

① holds due to Lemma 3, that is,

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

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

$$\begin{aligned}
&\mathcal{R}_T^{DOG}\\
=&\mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}}\sum_{t=1}^{T}\sum_{i=1}^{n} f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t}) - f_t(\mathbf{x}_t^*;\xi_{i,t})\\
\leq & 11nT\beta G\eta + \left(L+\frac{\eta L^2}{\beta}+6\eta L^2+\frac{\beta}{\eta}\right)\frac{\eta^2}{(1-\rho)^2}\left(6nTG+4\sum_{t=1}^{T}\sum_{i=1}^{n}\sigma_{i,t}^2\right)\\
&+\left(\frac{1}{\beta}+4\right)n\left(H_t(\bar{\mathbf{x}}_1)-H_t(\bar{\mathbf{x}}_{t+1})\right)+3nTLG\eta^2\left(\frac{1}{\beta}+4\right)\\
&+\frac{\beta(\eta+8nL\eta^2)+2nL\eta^2}{n\beta}\sum_{t=1}^{T}\sum_{i=1}^{n}\sigma_{i,t}^2+\frac{n}{2\eta}\left(4\sqrt{R}M+R\right)\\
=&\eta nTG\left(11\beta+\frac{6\eta}{(1-\rho)^2}\left(L+\frac{\eta L^2}{\beta}+6\eta L^2+\frac{\beta}{\eta}\right)+3L\eta\left(\frac{1}{\beta}+4\right)\right)+\left(\frac{1}{\beta}+4\right)n\left(H_t(\bar{\mathbf{x}}_1)-H_t(\bar{\mathbf{x}}_{t+1})\right)\\
&+\left(\left(L+\frac{\eta L^2}{\beta}+6\eta L^2+\frac{\beta}{\eta}\right)\frac{4\eta^2}{(1-\rho)^2}+\frac{\beta(\eta+8nL\eta^2)+2nL\eta^2}{n\beta}\right)\sum_{t=1}^{T}\sum_{i=1}^{n}\sigma_{i,t}^2+\frac{n}{2\eta}\left(4\sqrt{R}M+R\right).
\end{aligned}$$

It completes the proof. $\square$

Lemma 1. *For any β with $0 \le \beta \le 1$, we have*

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

Proof.

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

$\square$

Lemma 2.

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

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

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

① holds due to Lemma 1. Finally, we obtain

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

It completes the proof.

□

Lemma 3.

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

Proof. Recall that

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

and

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

Denote

$$
\mathbf{X}_t = [\mathbf{x}_{1,t}, \mathbf{x}_{2,t}, ..., \mathbf{x}_{n,t}] \in \mathbb{R}^{d\times n},
$$

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

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

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

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 4.

Thus, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\Xi_{n,T}\sim\mathcal{D}_{n,T}} \sum_{i=1}^{n}\sum_{t=1}^{T} \left\|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\right\|^2 \\
\le& \mathop{\mathbb{E}}_{\Xi_{n,T}\sim\mathcal{D}_{n,T}} \sum_{t=1}^{T} \left(\sum_{s=1}^{t-1} \eta\rho^{t-s-1}\left\|\mathbf{G}_s\right\|_F\right)^2 \\
\overset{①}{\le}& \frac{\eta^2}{(1-\rho)^2} \mathop{\mathbb{E}}_{\Xi_{n,T}\sim\mathcal{D}_{n,T}} \left(\sum_{t=1}^{T} \left\|\mathbf{G}_t\right\|_F^2\right) \\
=& \frac{\eta^2}{(1-\rho)^2} \left(\mathop{\mathbb{E}}_{\Xi_{n,T}\sim\mathcal{D}_{n,T}} \sum_{t=1}^{T}\sum_{i=1}^{n} \left\|\nabla f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t})\right\|^2\right) \\
=& \frac{\eta^2}{(1-\rho)^2} \left(\sum_{t=1}^{T}\sum_{i=1}^{n} \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \left\|\nabla f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t})\right\|^2\right) \\
\overset{②}{\le}& \frac{\eta^2}{(1-\rho)^2} \left(6nTG + 4\sum_{t=1}^{T}\sum_{i=1}^{n} \sigma_{i,t}^2\right).
\end{aligned}$$

① holds due to Lemma 5. ② holds due to Lemma 1. It completes the proof.

□

Lemma 4 (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 5 (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.