

# Decentralized Online Optimization

December 14, 2018

**Abstract**

ddd

## 1 Notations and assumptions

Define the dynamic regret as

$$\mathcal{R}_T^{DOG} = \mathbb{E} \sum_{\xi} \sum_{i=1}^n \sum_{t=1}^T f_{i,t}(\mathbf{x}_{i,t}; \xi) - f_{i,t}(\mathbf{x}_t^*; \xi),$$

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

$$f_{i,t}(\mathbf{x}; \xi) := \beta \bar{f}_{i,t}(\mathbf{x}) + (1 - \beta) f(\mathbf{x}; \xi)$$

with  $0 < \beta < 1$ , and  $\xi$  is a random variable drawn from an unknown distribution.  $\bar{f}_{i,t}$  is an adversary loss function, and  $f$  is a given loss function.

The budget of the dynamics is defined as

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

**Assumption 1.** For any  $i \in [n]$  and  $t \in [T]$ , we assume  $\|\nabla f_{i,t}(\mathbf{x})\|^2 \leq G$ . For any  $\mathbf{x} \in \mathcal{X}$  and  $\mathbf{y} \in \mathcal{X}$ , we assume  $\|\mathbf{x} - \mathbf{y}\|^2 \leq R$ .

**Assumption 2.** For any  $i \in [n]$  and  $t \in [T]$ , we assume the function  $f_{i,t}(\mathbf{x})$  is differentiable with respect to any vector  $\mathbf{x} \in \mathcal{X}$ .

## 2 Algorithm

**Theorem 1.** Using Assumptions 1 and 2, and choosing  $\eta > 0$  in Algorithm 1, we have

$$\begin{aligned} \mathcal{R}_T^{DOG} &= \sum_{i=1}^n \sum_{t=1}^T f_{i,t}(\mathbf{x}_{i,t}) - f_t(\mathbf{x}_t^*) \\ &\leq 13nT\beta G\eta + \left( (1 - \beta) \left( L + \frac{\eta L^2}{\beta} + 3\eta L^2 \right) + \frac{\beta}{\eta} \right) \frac{nTG\eta^2}{(1 - \rho)^2} \\ &\quad + n \left( \frac{1}{\beta} + 2 \right) (f(\bar{\mathbf{x}}_1) - f(\bar{\mathbf{x}}_{T+1})) + \frac{n}{2\eta} (4\sqrt{RD} + R). \end{aligned}$$

**Corollary 1.** Using Assumptions 1 and 2, and choosing  $\eta = \sqrt{\frac{D}{\beta T}}$  in Algorithm 1, we have

$$\mathcal{R}_T^{DOG} \leq \mathcal{O} \left( n\sqrt{\beta TD} \right).$$

---

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

---

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

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

## Appendix

### Proof details of Theorem 1:

*Proof.*

$$\begin{aligned} & \mathbb{E} \frac{1}{n} \sum_{i=1}^n f_{i,t}(\mathbf{x}_{i,t}) - f_t(\mathbf{x}_t^*) \\ &= \frac{1}{n} \sum_{i=1}^n \beta (\bar{f}_{i,t}(\mathbf{x}_{i,t}) - f_t(\mathbf{x}_t^*)) + (1 - \beta) \mathbb{E} (f(\mathbf{x}_{i,t}; \xi) - f(\mathbf{x}_t^*)) \\ &= \frac{1}{n} \sum_{i=1}^n \beta (\bar{f}_{i,t}(\mathbf{x}_{i,t}) - f_t(\mathbf{x}_t^*)) + (1 - \beta) (f(\mathbf{x}_{i,t}) - f(\mathbf{x}_t^*)) \\ &\leq \frac{1}{n} \sum_{i=1}^n \beta \langle \nabla \bar{f}_{i,t}(\mathbf{x}_{i,t}), \mathbf{x}_{i,t} - \mathbf{x}_t^* \rangle + (1 - \beta) \langle \nabla f(\mathbf{x}_{i,t}), \mathbf{x}_{i,t} - \mathbf{x}_t^* \rangle \\ &= \frac{1}{n} \sum_{i=1}^n \beta (\langle \nabla \bar{f}_{i,t}(\mathbf{x}_{i,t}), \mathbf{x}_{i,t} - \bar{\mathbf{x}}_t \rangle + \langle \nabla \bar{f}_{i,t}(\mathbf{x}_{i,t}), \bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1} \rangle + \langle \nabla \bar{f}_{i,t}(\mathbf{x}_{i,t}), \bar{\mathbf{x}}_{t+1} - \mathbf{x}_t^* \rangle) \\ &\quad + \frac{1}{n} \sum_{i=1}^n (1 - \beta) (\langle \nabla f(\mathbf{x}_{i,t}), \mathbf{x}_{i,t} - \bar{\mathbf{x}}_t \rangle + \langle \nabla f(\mathbf{x}_{i,t}), \bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1} \rangle + \langle \nabla f(\mathbf{x}_{i,t}), \bar{\mathbf{x}}_{t+1} - \mathbf{x}_t^* \rangle) \\ &= \frac{1}{n} \sum_{i=1}^n \beta (\underbrace{\langle \nabla \bar{f}_{i,t}(\mathbf{x}_{i,t}), \mathbf{x}_{i,t} - \bar{\mathbf{x}}_t \rangle + \langle \nabla \bar{f}_{i,t}(\mathbf{x}_{i,t}), \bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1} \rangle}_{I_1(t)}) \\ &\quad + \frac{1}{n} \sum_{i=1}^n (1 - \beta) (\underbrace{\langle \nabla f(\mathbf{x}_{i,t}), \mathbf{x}_{i,t} - \bar{\mathbf{x}}_t \rangle + \langle \nabla f(\mathbf{x}_{i,t}), \bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1} \rangle}_{I_2(t)}) \\ &\quad + \frac{1}{n} \sum_{i=1}^n \underbrace{\langle \nabla f_{i,t}(\mathbf{x}_{i,t}), \bar{\mathbf{x}}_{t+1} - \mathbf{x}_t^* \rangle}_{I_3(t)} \end{aligned}$$

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

$$I_1(t) \leq \frac{\beta}{n} \sum_{i=1}^n \left( \frac{\eta}{2} \|\nabla \bar{f}_{i,t}(\mathbf{x}_{i,t})\|^2 + \frac{1}{2\eta} \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + \frac{\eta}{2} \|\nabla f_{i,t}(\mathbf{x}_{i,t})\|^2 + \frac{1}{2\eta} \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2 \right)$$

$$\leq \beta G \eta + \frac{\beta}{2n\eta} \sum_{i=1}^n \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + \frac{\beta}{2n\eta} \sum_{i=1}^n \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2.$$

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

$$I_2(t) = (1 - \beta) \left( \underbrace{\frac{1}{n} \sum_{i=1}^n \langle \nabla f(\mathbf{x}_{i,t}), \mathbf{x}_{i,t} - \bar{\mathbf{x}}_t \rangle}_{J_1(t)} + \underbrace{\frac{1}{n} \sum_{i=1}^n \langle \nabla f(\mathbf{x}_{i,t}), \bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1} \rangle}_{J_2(t)} \right).$$

For  $J_1(t)$ , we have

$$\begin{aligned} J_1(t) &= \frac{1}{n} \sum_{i=1}^n \langle \nabla f(\mathbf{x}_{i,t}), \mathbf{x}_{i,t} - \bar{\mathbf{x}}_t \rangle \\ &= \frac{1}{n} \sum_{i=1}^n \langle \nabla f(\mathbf{x}_{i,t}) - \nabla f(\bar{\mathbf{x}}_t), \mathbf{x}_{i,t} - \bar{\mathbf{x}}_t \rangle + \frac{1}{n} \sum_{i=1}^n \langle \nabla f(\bar{\mathbf{x}}_t), \mathbf{x}_{i,t} - \bar{\mathbf{x}}_t \rangle \\ &\leq \frac{L}{n} \sum_{i=1}^n \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + \frac{1}{n} \sum_{i=1}^n \langle \nabla f(\bar{\mathbf{x}}_t), \mathbf{x}_{i,t} - \bar{\mathbf{x}}_t \rangle \\ &\leq \frac{L}{n} \sum_{i=1}^n \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + \frac{1}{n} \sum_{i=1}^n \left( \frac{\eta}{2\beta} \|\nabla f(\bar{\mathbf{x}}_t)\|^2 + \frac{\beta}{2\eta} \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 \right). \end{aligned} \quad (1)$$

According to Lemma 1, we have

$$\begin{aligned} &\frac{\eta}{2} \|\nabla f(\bar{\mathbf{x}}_t)\|^2 \\ &\leq \frac{\eta}{2} \|\nabla f(\bar{\mathbf{x}}_t)\|^2 + \left( \frac{\eta}{2} - \frac{L\eta^2}{2} \right) \left\| \frac{1}{n} \sum_{i=1}^n f_{i,t}(\mathbf{x}_{i,t}) \right\|^2 \\ &\leq f(\bar{\mathbf{x}}_t) - f(\bar{\mathbf{x}}_{t+1}) + 4G\eta\beta^2 + \frac{\eta L^2(1-\beta)^2}{n} \sum_{i=1}^n \|\bar{\mathbf{x}}_t - \mathbf{x}_{i,t}\|^2. \end{aligned} \quad (2)$$

Substituting (2) into (1), we obtain

$$\begin{aligned} J_1(t) &\leq \frac{L}{n} \sum_{i=1}^n \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + \left( \frac{1}{\beta} (f(\bar{\mathbf{x}}_t) - f(\bar{\mathbf{x}}_{t+1})) + 4G\eta\beta + \frac{\eta L^2(1-\beta)^2}{n\beta} \sum_{i=1}^n \|\bar{\mathbf{x}}_t - \mathbf{x}_{i,t}\|^2 + \frac{\beta}{2n\eta} \sum_{i=1}^n \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 \right) \\ &= \left( \frac{L}{n} + \frac{\eta L^2(1-\beta)^2}{n\beta} + \frac{\beta}{2n\eta} \right) \sum_{i=1}^n \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + \frac{1}{\beta} (f(\bar{\mathbf{x}}_t) - f(\bar{\mathbf{x}}_{t+1})) + 4G\eta\beta. \end{aligned}$$

For  $J_2(t)$ , we have

$$\begin{aligned} J_2(t) &= \frac{1}{n} \sum_{i=1}^n \langle \nabla f(\mathbf{x}_{i,t}), \bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1} \rangle \\ &\leq \frac{1}{n} \sum_{i=1}^n \left( \frac{\eta}{2} \|\nabla f(\mathbf{x}_{i,t})\|^2 + \frac{1}{2\eta} \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2 \right) \\ &\leq \frac{1}{n} \sum_{i=1}^n \left( \frac{\eta}{2} \|\nabla f(\mathbf{x}_{i,t}) - \nabla f(\bar{\mathbf{x}}_t) + \nabla f(\bar{\mathbf{x}}_t)\|^2 + \frac{1}{2\eta} \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2 \right) \end{aligned}$$

$$\begin{aligned}
&\leq \frac{1}{n} \sum_{i=1}^n \left( \eta \|\nabla f(\mathbf{x}_{i,t}) - \nabla f(\bar{\mathbf{x}}_t)\|^2 + \eta \|\nabla f(\bar{\mathbf{x}}_t)\|^2 + \frac{1}{2\eta} \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2 \right) \\
&\leq \frac{1}{n} \sum_{i=1}^n \left( \eta L^2 \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + \eta \|\nabla f(\bar{\mathbf{x}}_t)\|^2 + \frac{1}{2\eta} \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2 \right).
\end{aligned}$$

Recall Lemma 1, and we have

$$\begin{aligned}
J_2(t) &= \frac{1}{n} \sum_{i=1}^n \left( \eta L^2 \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + \left( 2(f(\bar{\mathbf{x}}_t) - f(\bar{\mathbf{x}}_{t+1})) + 8G\eta\beta^2 + \frac{2\eta L^2(1-\beta)^2}{n} \sum_{i=1}^n \|\bar{\mathbf{x}}_t - \mathbf{x}_{i,t}\|^2 \right) \right) + \frac{1}{2\eta} \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2 \\
&= \frac{\eta L^2(1+2(1-\beta)^2)}{n} \sum_{i=1}^n \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + 2(f(\bar{\mathbf{x}}_t) - f(\bar{\mathbf{x}}_{t+1})) + 8G\eta\beta^2 + \frac{1}{2\eta} \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2 \\
&\leq \frac{3\eta L^2}{n} \sum_{i=1}^n \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + 2(f(\bar{\mathbf{x}}_t) - f(\bar{\mathbf{x}}_{t+1})) + 8G\eta\beta^2 + \frac{1}{2\eta} \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2.
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
I_2(t) &= (1-\beta)(J_1(t) + J_2(t)) \\
&\leq (1-\beta) \left( \left( \frac{L}{n} + \frac{\eta L^2(1-\beta)^2}{n\beta} + \frac{\beta}{2n\eta} + \frac{3\eta L^2}{n} \right) \sum_{i=1}^n \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + \left( \frac{1}{\beta} + 2 \right) (f(\bar{\mathbf{x}}_t) - f(\bar{\mathbf{x}}_{t+1})) \right) \\
&\quad + (1-\beta) \left( 4G\eta\beta(1+2\beta) + \frac{1}{2\eta} \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2 \right).
\end{aligned}$$

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

$$\begin{aligned}
&I_1(t) + I_2(t) \\
&\leq \beta G\eta + \frac{\beta}{2n\eta} \sum_{i=1}^n \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + \frac{\beta}{2n\eta} \sum_{i=1}^n \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2 \\
&\quad + (1-\beta) \left( \left( \frac{L}{n} + \frac{\eta L^2(1-\beta)^2}{n\beta} + \frac{\beta}{2n\eta} + \frac{3\eta L^2}{n} \right) \sum_{i=1}^n \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 + \left( \frac{1}{\beta} + 2 \right) (f(\bar{\mathbf{x}}_t) - f(\bar{\mathbf{x}}_{t+1})) \right) \\
&\quad + (1-\beta) \left( 4G\eta\beta(1+2\beta) + \frac{1}{2\eta} \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2 \right) \\
&= (1+4(1-\beta)(1+2\beta))\beta G\eta + \left( (1-\beta) \left( \frac{L}{n} + \frac{\eta L^2(1-\beta)^2}{n\beta} + \frac{\beta}{2n\eta} + \frac{3\eta L^2}{n} \right) + \frac{\beta}{2n\eta} \right) \sum_{i=1}^n \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 \\
&\quad + \left( \frac{1}{\beta} + 2 \right) (f(\bar{\mathbf{x}}_t) - f(\bar{\mathbf{x}}_{t+1})) + \frac{1}{2\eta} \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2 \\
&\leq 13\beta G\eta + \left( (1-\beta) \left( \frac{L}{n} + \frac{\eta L^2}{n\beta} + \frac{3\eta L^2}{n} \right) + \frac{\beta}{n\eta} \right) \sum_{i=1}^n \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 \\
&\quad + \left( \frac{1}{\beta} + 2 \right) (f(\bar{\mathbf{x}}_t) - f(\bar{\mathbf{x}}_{t+1})) + \frac{1}{2\eta} \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2.
\end{aligned}$$

According to Lemma 2, we have

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

Therefore,

$$\begin{aligned}
& \sum_{t=1}^T (I_1(t) + I_2(t)) \\
& \leq 13TG\eta + \left( (1-\beta) \left( \frac{L}{n} + \frac{\eta L^2}{n\beta} + \frac{3\eta L^2}{n} \right) + \frac{\beta}{n\eta} \right) \sum_{t=1}^T \sum_{i=1}^n \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 \\
& \quad + \left( \frac{1}{\beta} + 2 \right) \sum_{t=1}^T (f(\bar{\mathbf{x}}_t) - f(\bar{\mathbf{x}}_{t+1})) + \frac{1}{2\eta} \sum_{t=1}^T \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2 \\
& \leq 13TG\eta + \left( (1-\beta) \left( \frac{L}{n} + \frac{\eta L^2}{n\beta} + \frac{3\eta L^2}{n} \right) + \frac{\beta}{n\eta} \right) \frac{nTG\eta^2}{(1-\rho)^2} \\
& \quad + \left( \frac{1}{\beta} + 2 \right) \sum_{t=1}^T (f(\bar{\mathbf{x}}_1) - f(\bar{\mathbf{x}}_{T+1})) + \frac{1}{2\eta} \sum_{t=1}^T \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2.
\end{aligned}$$

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

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

By taking average over  $i \in [n]$  on both sides, we have

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

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

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

Note that (3) is equivalent to

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

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

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

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

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

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

Dividing  $\tau$  on both sides, and letting  $\tau$  be close to 0, we have

$$\begin{aligned} I_3(t) &= \left\langle \frac{1}{n} \sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t}), \bar{\mathbf{x}}_{t+1} - \mathbf{x}_t^* \right\rangle \\ &\leq \frac{1}{2\eta} \left( \lim_{\tau \rightarrow 0} \tau \|\mathbf{x}_t^* - \bar{\mathbf{x}}_{t+1}\|^2 + 2 \langle \mathbf{x}_t^* - \bar{\mathbf{x}}_{t+1}, \bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_t \rangle \right) \\ &= \frac{1}{\eta} \langle \mathbf{x}_t^* - \bar{\mathbf{x}}_{t+1}, \bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_t \rangle \\ &= \frac{1}{2\eta} \left( \|\mathbf{x}_t^* - \bar{\mathbf{x}}_t\|^2 - \|\mathbf{x}_t^* - \bar{\mathbf{x}}_{t+1}\|^2 - \|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\|^2 \right). \end{aligned} \quad (4)$$

Besides, we have

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

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

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

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

$$\begin{aligned} &\mathbb{E} \sum_{t=1}^T \sum_{i=1}^n f_{i,t}(\mathbf{x}_{i,t}) - f_t(\mathbf{x}_t^*) \\ &\leq n \sum_{t=1}^T (I_1(t) + I_2(t) + I_3(t)) \\ &\leq 13nT\beta G\eta + \left( (1-\beta) \left( L + \frac{\eta L^2}{\beta} + 3\eta L^2 \right) + \frac{\beta}{\eta} \right) \frac{nTG\eta^2}{(1-\rho)^2} \\ &\quad + n \left( \frac{1}{\beta} + 2 \right) (f(\bar{\mathbf{x}}_1) - f(\bar{\mathbf{x}}_{T+1})) + \frac{n}{2\eta} (4\sqrt{R}D + R). \end{aligned}$$

□

**Lemma 1.**

$$\frac{\eta}{2} \|\nabla f(\bar{\mathbf{x}}_t)\|^2 + \left( \frac{\eta}{2} - \frac{L\eta^2}{2} \right) \left\| \frac{1}{n} \sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t}) \right\|^2 \leq f(\bar{\mathbf{x}}_t) - f(\bar{\mathbf{x}}_{t+1}) + 4G\eta\beta^2 + \frac{\eta L^2(1-\beta)^2}{n} \sum_{i=1}^n \|\bar{\mathbf{x}}_t - \mathbf{x}_{i,t}\|^2.$$

*Proof.*

$$\begin{aligned}
f(\bar{\mathbf{x}}_{t+1}) &\leq f(\bar{\mathbf{x}}_t) + \langle \nabla f(\bar{\mathbf{x}}_t), \bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_t \rangle + \frac{L}{2} \|\bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_t\|^2 \\
&= f(\bar{\mathbf{x}}_t) + \left\langle \nabla f(\bar{\mathbf{x}}_t), -\frac{\eta}{n} \sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t}) \right\rangle + \frac{L}{2} \left\| \frac{\eta}{n} \sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t}) \right\|^2 \\
&= f(\bar{\mathbf{x}}_t) + \frac{\eta}{2} \left( \left\| \nabla f(\bar{\mathbf{x}}_t) - \frac{1}{n} \sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t}) \right\|^2 - \|\nabla f(\bar{\mathbf{x}}_t)\|^2 - \left\| \frac{1}{n} \sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t}) \right\|^2 \right) + \frac{L}{2} \left\| \frac{\eta}{n} \sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t}) \right\|^2 \\
&= f(\bar{\mathbf{x}}_t) + \frac{\eta}{2} \left\| \nabla f(\bar{\mathbf{x}}_t) - \frac{1}{n} \sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t}) \right\|^2 - \frac{\eta}{2} \|\nabla f(\bar{\mathbf{x}}_t)\|^2 - \left( \frac{\eta}{2} - \frac{L\eta^2}{2} \right) \left\| \frac{1}{n} \sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t}) \right\|^2. \quad (4)
\end{aligned}$$

Additionally, we have

$$\begin{aligned}
&\left\| \nabla f(\bar{\mathbf{x}}_t) - \frac{1}{n} \sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t}) \right\|^2 \\
&= \left\| \nabla f(\bar{\mathbf{x}}_t) - \frac{1}{n} \sum_{i=1}^n (\beta \nabla \bar{f}_{i,t}(\mathbf{x}_{i,t}) + (1-\beta) \nabla f(\mathbf{x}_{i,t})) \right\|^2 \\
&\leq 2\beta^2 \left\| \nabla f(\bar{\mathbf{x}}_t) - \frac{1}{n} \sum_{i=1}^n \nabla \bar{f}_{i,t}(\mathbf{x}_{i,t}) \right\|^2 + 2(1-\beta)^2 \left\| \nabla f(\bar{\mathbf{x}}_t) - \frac{1}{n} \sum_{i=1}^n \nabla f(\mathbf{x}_{i,t}) \right\|^2 \\
&\leq 2\beta^2 \left( 2\|\nabla f(\bar{\mathbf{x}}_t)\|^2 + 2 \left\| \frac{1}{n} \sum_{i=1}^n \nabla \bar{f}_{i,t}(\mathbf{x}_{i,t}) \right\|^2 \right) + 2(1-\beta)^2 \left\| \nabla f(\bar{\mathbf{x}}_t) - \frac{1}{n} \sum_{i=1}^n \nabla f(\mathbf{x}_{i,t}) \right\|^2 \\
&\leq 8G\beta^2 + 2(1-\beta)^2 \left\| \nabla f(\bar{\mathbf{x}}_t) - \frac{1}{n} \sum_{i=1}^n \nabla f(\mathbf{x}_{i,t}) \right\|^2 \\
&\leq 8G\beta^2 + \frac{2(1-\beta)^2}{n} \sum_{i=1}^n \|\nabla f(\bar{\mathbf{x}}_t) - \nabla f(\mathbf{x}_{i,t})\|^2 \\
&\leq 8G\beta^2 + \frac{2L^2(1-\beta)^2}{n} \sum_{i=1}^n \|\bar{\mathbf{x}}_t - \mathbf{x}_{i,t}\|^2. \quad (5)
\end{aligned}$$

Substituting (5) into (4), we obtain

$$f(\bar{\mathbf{x}}_{t+1}) \leq f(\bar{\mathbf{x}}_t) + \frac{\eta}{2} \left( 8G\beta^2 + \frac{2L^2(1-\beta)^2}{n} \sum_{i=1}^n \|\bar{\mathbf{x}}_t - \mathbf{x}_{i,t}\|^2 \right) - \frac{\eta}{2} \|\nabla f(\bar{\mathbf{x}}_t)\|^2 - \left( \frac{\eta}{2} - \frac{L\eta^2}{2} \right) \left\| \frac{1}{n} \sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t}) \right\|^2.$$

Equivalently, we obtain

$$\frac{\eta}{2} \|\nabla f(\bar{\mathbf{x}}_t)\|^2 + \left( \frac{\eta}{2} - \frac{L\eta^2}{2} \right) \left\| \frac{1}{n} \sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t}) \right\|^2 \leq f(\bar{\mathbf{x}}_t) - f(\bar{\mathbf{x}}_{t+1}) + 4G\eta\beta^2 + \frac{\eta L^2(1-\beta)^2}{n} \sum_{i=1}^n \|\bar{\mathbf{x}}_t - \mathbf{x}_{i,t}\|^2.$$

It completes the proof.  $\square$

**Lemma 2.**

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

*Proof.* Recall that

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

and

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

Denote

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

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

$$\bar{\mathbf{x}}_{t+1} = \bar{\mathbf{x}}_t - \eta \left( \frac{1}{n} \sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t}) \right) = - \sum_{s=1}^t \eta \bar{\mathbf{G}}_s. \quad (6)$$

Therefore,

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

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

According to Lemma 4, we have

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \sum_{t=1}^T \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2 & \leq \frac{1}{n(1-\rho)^2} \sum_{t=1}^T \eta^2 \|\mathbf{G}_t\|_F^2 \\ & \leq \frac{1}{(1-\rho)^2} \sum_{t=1}^T \eta^2 G. \end{aligned}$$

It completes the proof. □



**Lemma 3** (Appeared in Lemma 5 in [Tang et al., 2018]). For any matrix  $\mathbf{X}_t \in \mathbb{R}^{d \times n}$ , decompose the confusion matrix  $\mathbf{W}$  as  $\mathbf{W} = \sum_{i=1}^n \lambda_i \mathbf{v}_i \mathbf{v}_i^\top = \mathbf{P} \mathbf{\Lambda} \mathbf{P}^\top$ , where  $\mathbf{P} = [\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n] \in \mathbb{R}^{n \times n}$ ,  $\mathbf{v}_i$  is the normalized eigenvector of  $\lambda_i$ .  $\mathbf{\Lambda}$  is a diagonal matrix, and  $\lambda_i$  be its  $i$ -th element. We have

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

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

**Lemma 4** (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

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

## References

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