Gossip Online Learning: Exchanging Local Models to Track Dynamics

January 19, 2019

Abstract

In this paper, we consider online learning in the decentralized setting up, which is motivated by the scenario where users want to take benefits from the data from other users, but do not want to share their private data to a third party or other users. Instead, they can only share their private prediction model, e.g., recommendation model. We study the decentralized online gradient method in which each user maintains a private model and share its private model with its neighbors (or users he/she trusts) periodically. In addition, to consider more practical scenario we allow users' interest changing over time, unlike most online work which assumes that the optimal prediction model is constant. We prove that decentralized online gradient can efficiently and effectively propagate the values in all private data without leaking them to track the dynamics, by admitting a tight dynamic regret $\mathcal{O}\left(n\sqrt{TM}+\sqrt{nTM}\sigma\right)$ where n is the number of users, T is the number of time steps, M measures the dynamics (this is, how much the users' interest changes over time), and σ measures the randomness of the private data. Empirical studies are also conducted to validate our analysis.

1 Introduction

Online learning has been studied for decades of years in machine learning literatures [Hazan, 2016, Shalev-Shwartz, 2012, Duchi et al., 2011]. The goal of online learning generally is to incrementally learn predictions models to minimize the sum of all the online loss functions (cumulative loss), which is usually determined by a squence of examples that arrives sequentially. To quantify the efficacy of an online learning algorithm, the community introduced a performance measure called *static regret*, which is the difference between the cumulative losses suffered by the online algorithm and that suffered by the best model which can observe all the loss functions. The best static regret of a sequential online convex optimization method is $\mathcal{O}\left(\sqrt{T}\right)$ and $\mathcal{O}\left(\log T\right)$ for convex and strongly convex loss functions, respectively [Hazan, 2016, Shalev-Shwartz, 2012].

Different with traditional online learning, online learning in decentralized networks (or Decentralized Online Learning) assumes that a network of computational nodes can communicate between neighors to solve an online learning problem, in which each computational node will receive a stream of online losses. Suppose we have n workers, among which the i-th one will receive the t-th loss $f_{i,t}$ at the $t-$th iteration. Then, the goal of Decentralized Online Learning usually is to minimize its static regret, which is defined as the difference between the cumulative loss over all the nodes and steps and that of the best model which knows all the loss function beforehand. Decentralized Online Learning enjoys many advantages for real-world large-scale applications. Firstly, it avoid collecting all the loss functions to one node, which will result in heavy communication cost for the network and extremely high computational cost for one node. Secondly, it can help many data providers collaborate to better minimize their cumulative loss, while at the same time protect the data privacy as much as possible.

The static regret assumes that the best model keeps unchanged during the entire learning process, however this does not hold in some real applications. For example, one's favorite style of musics may change over time as his/her situation. To solve this issue, the dynamic regret is introduced, which generally measure the difference between the cumulative loss suffered by the decentralized online learning algorithm and that suffered by a dynamic sequence of models. This dynamic sequence of models can not only observe all the loss functions beforehand, but also changes over time with the amount of changes less than a budget. In this paper, we mainly prove that decentralized online gradient can achieve a dynamic regret of $\mathcal{O}\left(n\sqrt{TM}+\sqrt{nTM}\sigma\right)$ where n is the number of users, T is the number of time steps, M measures the dynamics budget, and σ measures the randomness of the private data.

Notations and definitions In the paper, we make the following notations.

- For any $i \in [n]$ and $t \in [T]$, the random variable $\xi_{i,t}$ is subject to a distribution D_t, that is, $\xi_{i,t} \sim D_t$. Besides, a set of random variables $\Xi_{n,T}$ and the corresponding set of distributions are defined by

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

  respectively. For math brevity, we use the notation $\Xi_{n,T} \sim \mathcal{D}_T$ to represent that $\xi_{i,t} \sim D_t$ holds for any $i \in [n]$ and $t \in [T]$. $\mathbb{E}$ represents mathematical expectation.

- For a decentralized network, we use $\mathbf{W} \in \mathbb{R}^{n\times n}$ to represent its confusion matrix. It is a symmetric doubly stochastic matrix, which implies that every element of $\mathbf{W}$ is non-negative, $\mathbf{W1} = \mathbf{1}$, and $\mathbf{1}^{\mathrm{T}}\mathbf{W} = \mathbf{1}^{\mathrm{T}}$. We use $\{\lambda_i\}_{i=1}^n$ with $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_n$ to represent its eigenvalues.

- ∂ and ∇ represent sub-gradient and gradient operators, respectively. $\|\cdot\|$ represents the ℓ_2 norm in default.

- $\lesssim$ represents "less than equal up to a constant factor".

2 Related work

Online learning has been studied for decades of years. The static regret of a sequential online convex optimization method can achieve $\mathcal{O}\left(\sqrt{T}\right)$ and $\mathcal{O}(\log T)$ bounds for convex and strongly convex loss functions, respectively [Hazan, 2016, Shalev-Shwartz, 2012]. Recently, both the decentralized online learing and the dynamic regret have drawn much attention due to their wide existence in the practical big data scenarios.

2.1 Decentralized online learning

Online learning in a decentralized network has been studied in [Shahrampour and Jadbabaie, 2018, Kamp et al., 2014, Koppel et al., 2018, Zhang et al., 2018a, 2017b, Xu et al., 2015, Akbari et al., 2017, Lee et al., 2016, Nedi et al., 2015, Lee et al., 2018, Benczúr et al., 2018, Yan et al., 2013]. Shahrampour and Jadbabaie [2018] studies decentralized online mirror descent, and provides $\mathcal{O}\left(n\sqrt{nTM}\right)$ dynamic regret. Here, n, T, and M represent the number of nodes in the newtork, the number of iterations, and the budget of dynamics (defined in (2)), respectively. When the Bregman divergence in the decentralized online mirror descent is chosen appropriately, the decentralized online mirror descent becomes identical to the decentralized online gradient descent. Using the same definition of dynamic regret (defined in (3)), our method obtains $\mathcal{O}\left(n\sqrt{TM}\right)$ dynamic regret for a decentralized online gradient descent, which is better than $\mathcal{O}\left(n\sqrt{nTM}\right)$ in Shahrampour and Jadbabaie [2018]. The improvement of our bound benefits from a better bound of network error (see Lemma 1). Kamp et al. [2014] studies decentralized online prediction, and presents $\mathcal{O}\left(\sqrt{nT}\right)$ static regret. It assumes that all data, used to yielded the loss, is generated from an unknown distribution. The strong assumption is not practical in the dynamic environment, and thus limits its novelity for a general online learning task. Additionally, many decentralized online optimization methods are proposed, for example, decentralized online multi-task learning [Zhang et al., 2018a], decentralized online ADMM [Xu et al., 2015], decentralized online sub-gradient descent [Akbari et al., 2017], decentralized continuous-time online saddle-point method [Lee et al., 2016], decentralized online Nesterov's primal-dual method [Nedi et al., 2015, Lee et al., 2018]. Those previous methods are proved to yield $\mathcal{O}\left(\sqrt{T}\right)$ static regret, which do not have theoretical guarantee of regret in the dynamic environment. Besides, Yan et al. [2013] provides necessary and sufficient conditions to preserve privacy for decentralized online learning methods, which is interesting to extend our method to be privacy-preserving in the future work.

2.2 Regret in dynamic environment

Dynamic regret has been widely studied for decades of years [Zinkevich, 2003, Hall and Willett, 2015, 2013, Jadbabaie et al., 2015, Yang et al., 2016, Bedi et al., 2018, Zhang et al., 2017a, Mokhtari et al., 2016, Zhang et al., 2018b, György and Szepesvári, 2016, Wei et al., 2016, Zhao et al., 2018]. Zinkevich [2003] first defines the dynamic regret by (3), and then proposes an online gradient descent method. The method yields $\mathcal{O}\left(\sqrt{TM}\right)$ by choosing an appropriate learning rate. The following researches achieve the sublinear dynamic regret, but extend the analysis of regret by using different reference points. For example, Hall and Willett [2015, 2013] choose the reference points $\{\mathbf{x}_t^*\}_{t=1}^T$ satisfying $\sum_{t=1}^{T-1}\left\|\mathbf{x}_{t+1}^* - \Phi(\mathbf{x}_t^*)\right\| \leq M$, where $\Phi(\mathbf{x}_t^*)$ is the predictive optimal modelmodel. When the function Φ predicts accurately, a small M is enough to bound the dynamics. The dynamic regret is thus effectively decreased. Jadbabaie et al. [2015], Yang et al. [2016], Bedi et al. [2018], Zhang et al. [2017a], Mokhtari et al. [2016], Zhang et al. [2018b] chooses the reference points $\{\mathbf{y}_t^*\}_{t=1}^T$ with $\mathbf{y}_t^* = \operatorname{argmin}_{\mathbf{z}\in\mathcal{X}} f_t(\mathbf{z})$, where f_t is the loss function at the t-th iteration. György and Szepesvári [2016] provides a new analysis framework, which achieves $\mathcal{O}\left(\sqrt{TM}\right)$ dynamic regret[1] for any given reference points. Besides, Zhao et al. [2018] presents that the lower bound of the dynamic regret defined by 3 is $\Omega\left(\sqrt{TM}\right)$. The previous definition of the regret, i.e., (3), is a special case of our new definition. When setting $\beta = 1$, we achieve the state-of-the-art regret, that is, $\mathcal{O}\left(\sqrt{TM}\right)$.

In some literatures, the regret in a dynamic environment is measured by the number of changes of a reference point over time. It is usually denoted by shifting regret or tracking regret [Herbster and Warmuth, 1998, György et al., 2005, Gyorgy et al., 2012, György and Szepesvári, 2016, Mourtada and Maillard, 2017, Adamskiy et al., 2016, Wei et al., 2016, Cesa-Bianchi et al., 2012, Mohri and Yang, 2018, Jun et al., 2017]. Both the shifting regret and the tracking regret can be considered as a variation of the dynamic regret, and is usually studied in the setting of "learning with expert advice". But, the dynamic regret is usually studied in a general setting of online learning.

3 Problem formulation

For any a decentralized online algorithm $A \in \mathcal{A}$, we define its dynamic regret $\mathcal{R}_T^A$ by

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

where n is the number of nodes in the decentralized network. $\{\mathbf{x}_t^*\}_{t=1}^T$ is the sequence of reference points. $\mathbf{x}_{i,t}$ is the model played by an online algorithm A at the t-th round. The local loss function $f_{i,t}(\mathbf{x};\zeta_{i,t},\xi_{i,t})$ is defined by

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

[1] György and Szepesvári [2016] uses the notation of "shifting regret" instead of "dynamic regret". In the paper, we keep using "dynamic regret" as used in most previous literatures.

with $0 < \beta < 1$. $\zeta_{i,t}$ represents the adversary part of data. $\xi_{i,t}$ represents the stochastic part of data, which is drawn from the distribution D_t. Note that $g_{i,t}$ is an adversary loss function, which is caused by the adversary data. $h_t(\cdot;\xi_{i,t})$ is a stochastic loss function, which depends on the stochastic data $\xi_{i,t}$. The expectation is taken with respect to $\{\xi_{i,t}\}_{1\le i\le n, 1\le t\le T}$.

The sequence of reference points $\{\mathbf{x}_t^*\}_{t=1}^T$ satisfies

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

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

$$\sum_{t=1}^{T-1} \left\| \mathbf{x}_{t+1}^* - \mathbf{x}_t^* \right\| \le M. \tag{2}$$

When $M = 0$, all $\mathbf{x}_t^*$s are same, and it degenerates to the static online learning problem. When the dynamic environment changes significantly, M becomes large to model the dynamics. Let us take an example to explain the dynamics. Suppose we want to conduct online music recommendation task by using users' browsing records in Youtube. Every user has his/her own favorite music, and users' preference changes over time due to time-varying trends of hot topics in Internet. It leads to the dynamics of the optimal recommendation model.

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

$$\widetilde{\mathcal{R}}_T^A = \sum_{i=1}^n \sum_{t=1}^T \left(f_{i,t}(\mathbf{x}_{i,t}; \zeta_{i,t}, \xi_{i,t}) - f_{i,t}(\mathbf{x}_t^*) \right), \tag{3}$$

subject to $\sum_{t=1}^{T-1} \left\| \mathbf{x}_{t+1}^* - \mathbf{x}_t^* \right\| \le M$. In (3), the classic online learning in a decentralized network treats all the data as the adversary data. It ignores the potential relation of data among different nodes. Comparing with it, our definition of the dynamic regret, i.e., (1), views the adversary part of data and the stochastic part of data, distinctively. Since every node shares its private model to neighbours, the regret due to stochastic part of data would be decreased effectively, which is varified by the theoretical results in Section 4.2.

4 Decentralized online gradient method

In the section, we first present the decentralized online gradient method, and then prove that it leads to $\mathcal{O}\left(n\sqrt{TM} + \sqrt{nTM}\sigma\right)$ dynamic regret.

4.1 Algorithm

Algorithm 1 DOG: Decentralized Online Gradient method.

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

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

The Decentralized Online Gradient method, namely DOG, is presented in Algorithm 1. This algorithm works iteration by iteration. At each iteration, every node needs to collect local models, e.g., $\mathbf{x}_{i,t}$, from its neighbours, and compute a weighted sum as $\sum_{j=1}^{n} \mathbf{W}_{i,j}\mathbf{x}_{j,t}$. Then, the weighte sum is updated by an online gradient descent step. In addition, we denote $\bar{\mathbf{x}}_t = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_{i,t}$ to facilitate the theoretical analysis. We can verify that $\bar{\mathbf{x}}_{t+1} = \bar{\mathbf{x}}_t - \frac{\eta}{n}\sum_{i=1}^{n} \partial f_{i,t}(\mathbf{x}_{i,t};\zeta_{i,t},\xi_{i,t})$ (see Lemma 5).

4.2 Theoretical analysis

Denote

$$H_t(\cdot) := \mathop{\mathbb{E}}_{\xi_{i,t}\sim D_t} h_t(\cdot;\xi_{i,t}) \qquad \text{for } \forall i \in [n],$$

and

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

Assumption 1. *We make following assumptions to analyze the dynamic regret theoretically.*

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

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

  and

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

- *For given vectors $\mathbf{x}$ 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, but may be non-smooth.*
- *Given a symmetric doublely stochastic matrix $\mathbf{W}$, and a constant ρ with $\rho := \max\{|\lambda_2(\mathbf{W})|, |\lambda_n(\mathbf{W})|\}$, we assume $\rho < 1$.*

The bound of dynamic regret yielded by Algorithm 1 is presented in the following theorem.

Theorem 1. *Denote constants C_0, and C_1 by*

$$\begin{aligned} C_0 :=& 1 + \frac{1}{2(1-\rho)^2} + 16\beta; \\ C_1 :=& \frac{L + 2\eta L^2 + 4(1-\beta)^2 L^2\eta}{(1-\rho)^2} + 2L. \end{aligned}$$

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

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

By choosing an approximate learning rate η, we obtain sublinear regret as follows.

Corollary 1. *Using Assumption 1, and choosing*

$$\eta = \sqrt{\frac{(1-\rho)^2\left(nM\sqrt{R}+nR\right)}{nTG+(1-\beta)(1-\rho)^2T\sigma^2}}$$

in Algorithm 1, we have

$$\begin{aligned}
&\mathcal{R}_T^{\mathrm{DOG}} \\
\lesssim& \frac{n\sqrt{T\left(M+\sqrt{R}\right)G}}{1-\rho} + \sqrt{n(1-\beta)T\left(M+\sqrt{R}\right)\sigma^2} \\
&+ n\left(M+\sqrt{R}\right) + \frac{\sqrt{TM\left(n^2G+n(1-\beta)(1-\rho)^2\sigma^2\right)}}{1-\rho} \\
&+ \frac{\sqrt{T\left(n^2G+n(1-\beta)(1-\rho)^2\sigma^2\right)}}{1-\rho}.
\end{aligned} \tag{4}$$

First, corollary 1 shows that the dynamic regret of DOG is sublinear. Second, we would like make some comments on the effects of different parameters on the dynamic regret. The regret becomes large with the increase of the budget of dynamics M. When $n=1$ and $\rho=0$, the dynamic regret is $\mathcal{O}\left(\sqrt{TM}+\sqrt{T}\right)$, which is tight in the case of $n=1$ [Zhao et al., 2018]. When $\beta<1$, the regret $\mathcal{R}_T^{\mathrm{DOG}}$ has $\sqrt{nTM\sigma^2}$ dependence on σ^2, instead of $\sqrt{n^2TM\sigma^2}$. It benefits from the communication among nodes in the decentralized setting. Since every node shares its model with its neighbours, the variance of the average of stochastic gradients $\frac{1}{n}\sum_{i=1}^n \nabla h_t(\mathbf{x}_{i,t};\xi_{i,t})$ is decreased to be $\frac{\sigma^2}{n}$, thus eventually reducing the regret caused by the stochastic part of data. Additionally, the regret is affected by the topology of the network, which is measured by ρ with $0\le\rho<1$. For a fully connected network[2], $\rho=0$, then the regret is better than those for other topologies.

4.3 Connections with previous work

Improvement of dependence on n. Shahrampour and Jadbabaie [2018] investigates the dynamic regret $\widetilde{\mathcal{R}}_T^{\mathrm{DOG}}$ by using DOG, and provide the following sublinear regret.

Theorem 2 (Implied by Theorem 3 and Corollary 4 in Shahrampour and Jadbabaie [2018])**.** *Use Assumption 1, and choose $\eta=\sqrt{\frac{(1-\rho)M}{T}}$ in Algorithm 1. The dynamic regret $\widetilde{\mathcal{R}}_T^{\mathrm{DOG}}$ is bounded by $\mathcal{O}\left(n^{\frac{3}{2}}\sqrt{\frac{MT}{1-\rho}}\right)$.*

As illustrated in theorem 2, Shahrampour and Jadbabaie [2018] has provided a $\mathcal{O}\left(n\sqrt{nTM}\right)$ regret for DOG by using the previous dynamic regret defined in (3). Compared with the result in Shahrampour and Jadbabaie [2018], our regret enjoys the state-of-the-art dependence on T and M, and meanwhile improves the dependence on n. This improvement is achieved by a better bound on the difference between $\mathbf{x}_{i,t}$ and $\bar{\mathbf{x}}_t$[3].

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

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

Actually, the previous dynamic regret (3) is a special case of our dynamic regret by setting $\beta=1$.

[2] When a network is fully connected, a decentralized method de-generates to a centralized method.
[3] Shahrampour and Jadbabaie [2018] denotes $\|\mathbf{x}_{i,t}-\bar{\mathbf{x}}_t\|$ by "network error".

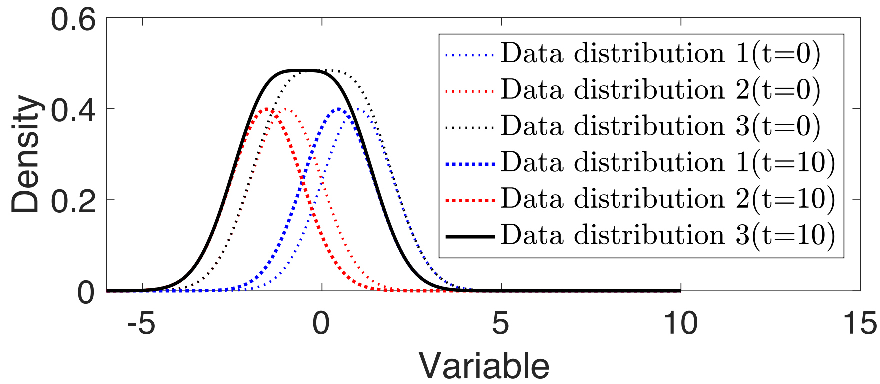


Figure 1: An illustration of the dynmaics caused by the time-varying distributions of data. Data distributions 1 and 2 satisify $N(1+\sin(t), 1)$ and $N(-1+\sin(t), 1)$, respectively. Data distribution 3 is the sum of them, which changes over time.

Improvement of dependence on σ^2. Previous researches [Shahrampour and Jadbabaie, 2018, Zhang et al., 2017b, Akbari et al., 2017] view all data as the adversary data, ignoring the potential relations among local models. They usually assume gradient of the loss function $\partial f_{i,t}$ is bounded, e.g., $\left\|\partial f_{i,t}(\mathbf{x};\zeta_{i,t},\xi_{i,t})\right\|^2 \leq G$, which implies $\left\|\nabla h_t(\mathbf{x};\xi_{i,t})\right\|^2 \leq G$, and $\mathbb{E}_{\xi_{i,t}\sim D_t}\left\|\nabla h_t(\mathbf{x};\xi_{i,t})\right\|^2 \leq \sigma^2 + G$ according to Lemma 2.

Lemma 2. *Assume* $\left\|\nabla h_t(\mathbf{x};\xi_{i,t})\right\|^2 \leq G$. *It implies*

$$\mathop{\mathbb{E}}_{\xi_{i,t}\sim D_t}\left\|\nabla h_t(\mathbf{x};\xi_{i,t})\right\|^2 \leq \sigma^2 + G.$$

Using this assumption in previous analysis frameworks, the regret $\mathcal{R}_T^{\mathrm{DOG}}$ has the same dependence on both G and σ^2 even in the static environment. However, our new analysis shows that the regret $\mathcal{R}_T^{\mathrm{DOG}}$ has $\sqrt{n\sigma^2}$ dependence on σ^2, and $\sqrt{n^2 G}$ dependence on G. The reason is that the variance of the average of stochastic gradients, i.e., $\nabla h_t(\cdot,\xi_{i,t})$ with $i \in [n]$, is decreased effectively when every node shares its local model to others.

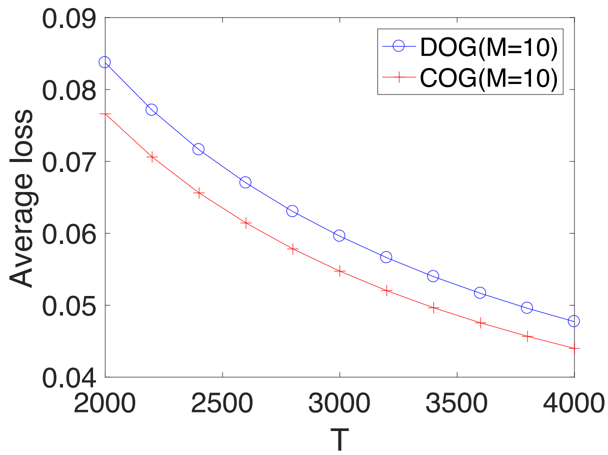
5 Empirical studies

For simplicity, in the experiments we only consider online logistic regression with squared ℓ_2 norm regularization, i.e., $f_{i,t}(\mathbf{x};\xi_{i,t}) = \log\left(1+\exp(-\mathbf{y}_{i,t}\mathbf{A}_{i,t}^{\mathrm{T}}\mathbf{x})\right) + \frac{\gamma}{2}\left\|\mathbf{x}\right\|^2$, where $\gamma = 10^{-3}$ is a given hyper-parameter. Under this setting, we compare the proposed Decentralized Online Gradient method (DOG) and the Centralized Online Gradient method (COG). The learning rate η is set to be $C\sqrt{\frac{M}{T}}$ with $C \in [10^{-2}, 20]$. M is fixed as 10 to determine the space of reference points, while C is tuned for each data sperately. We evaluate the learning performance by measuring the average loss $\frac{1}{nT}\sum_{i=1}^n\sum_{t=1}^T f_{i,t}(\mathbf{x}_{i,t};\zeta_{i,t},\xi_{i,t})$, instead of the dynamic regret $\mathbb{E}_{\Xi_{n,T}\sim\mathcal{D}_{n,T}}\sum_{i=1}^n\sum_{t=1}^T\left(f_{i,t}(\mathbf{x}_{i,t};\zeta_{i,t},\xi_{i,t}) - f_{i,t}(\mathbf{x}_t^*)\right)$, since the optimal reference point $\{\mathbf{x}_t^*\}_{t=1}^T$ is the same for DOG and COG .

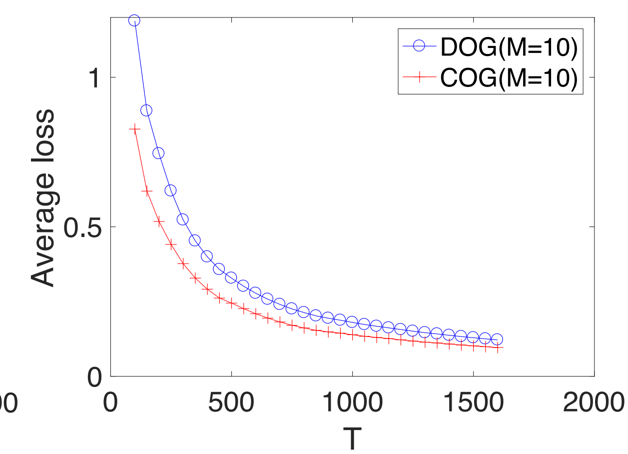
5.1 Datasets

To test the proposed algorithm, we utilized a toy dataset and three real-world datasets, whose details are as follows.

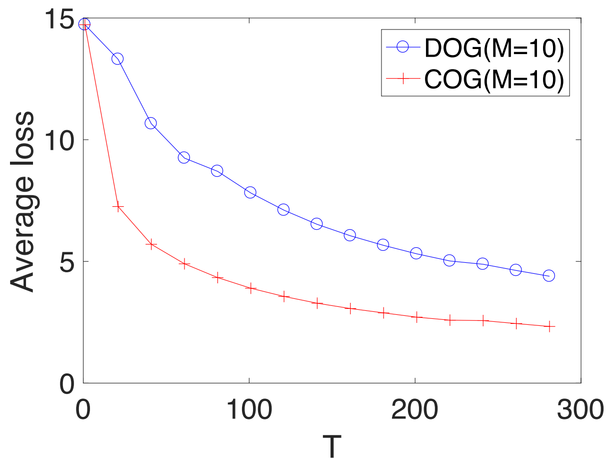
Synthetic Data For the i-th node, a data matrix $\mathbf{A}_i \in R^{10\times T}$ is generated, s.t. $\mathbf{A}_i = 0.1\tilde{\mathbf{A}}_i + 0.9\hat{\mathbf{A}}_i$, where $\tilde{\mathbf{A}}_i$ represents the adversary part of data, and $\hat{\mathbf{A}}_i$ represents the stochastic part of data. Specifically, elements of $\tilde{\mathbf{A}}_i$ is uniformly sampled from the interval $[-0.5+\sin(i), 0.5+\sin(i)]$. Note that $\tilde{\mathbf{A}}_i$ and $\tilde{\mathbf{A}}_j$



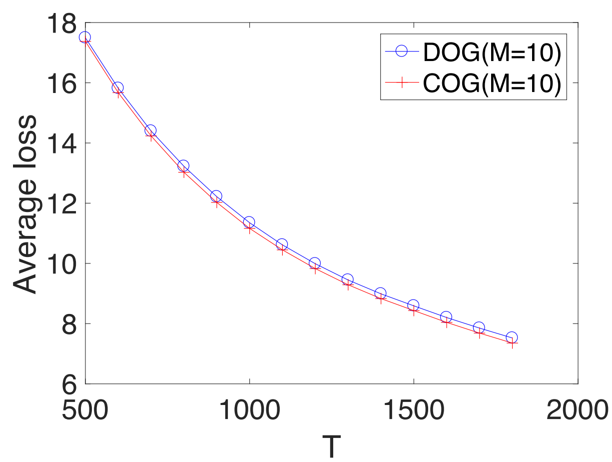
(a) *synthetic data*, 100 nodes



(b) *room-occupancy*, 5 nodes



(c) *usenet2*, 5 nodes



(d) *spam*, 5 nodes

Figure 2: The average loss yielded by DOG is comparable to that yielded by COG.

with $i \neq j$ are drawn from different distributions. $\hat{\mathbf{A}}_{i,t}$ is generated according to $\mathbf{y}_{i,t} \in \{1, -1\}$ which is generated uniformly. When $\mathbf{y}_{i,t} = 1$, $\hat{\mathbf{A}}_{i,t}$ is generated by sampling from a time-varying distribution $N((1+0.5\sin(t))\cdot\mathbf{1}, \mathbf{I})$. When $\mathbf{y}_{i,t} = -1$, $\hat{\mathbf{A}}_{i,t}$ is generated by sampling from another time-varying distribution $N((-1+0.5\sin(t))\cdot\mathbf{1}, \mathbf{I})$. Due to this correlation, $y_{i,t}$ can be considered as the label of the instance $\hat{\mathbf{A}}_{i,t}$. The above dynamics of time-varying distributions are illustrated in Figure 1, which shows the change of the optimal learning model over time and the importance of studying the dynamic regret.

Real Data Three real public datasets are *room-occupancy*[4], *usenet2*[5], and *spam*[6]. *room-occupancy* is a time-series dataset, which is from a natural dynamic enviroment. Both *usenet2* and *spam* are "concept drift" [Katakis et al., 2010] datasets, for which the optimal model changes over time.

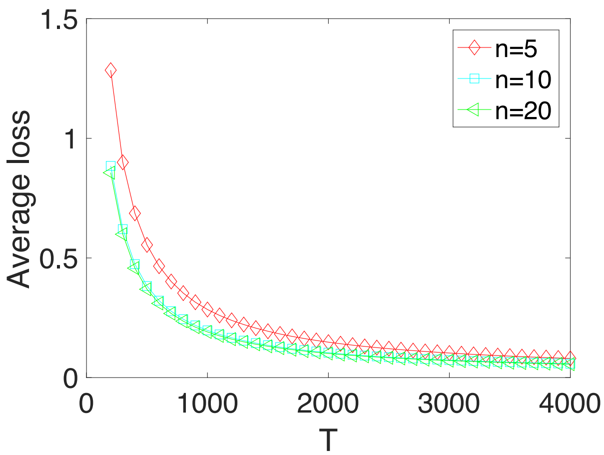
5.2 Results

First, figure 2 summarizes the performance of DOG compared with COG on all the datasets. For the synthetic dataset, we simulated a decentralized network consisting of 100 nodes; For the three real datasets,

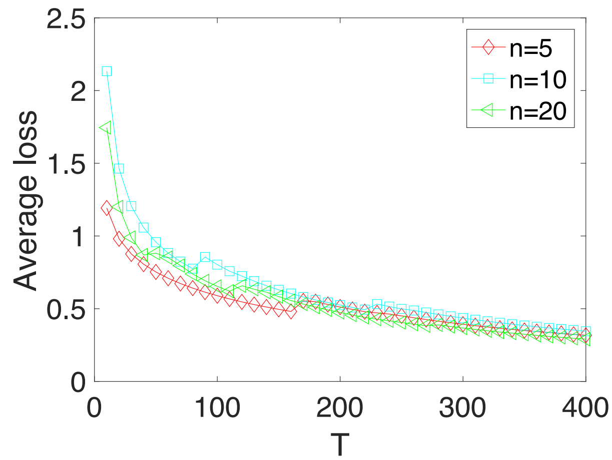
[4] https://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+
[5] http://mlkd.csd.auth.gr/concept_drift.html
[6] http://mlkd.csd.auth.gr/concept_drift.html

we simulated a network consisting of 5 nodes. In these networks, the nodes are connected by a ring topology. Under these settings, we can observe that both DOG and COG are effective for the online learning tasks on all the datasets, while DOG achieves slightly worse performance.

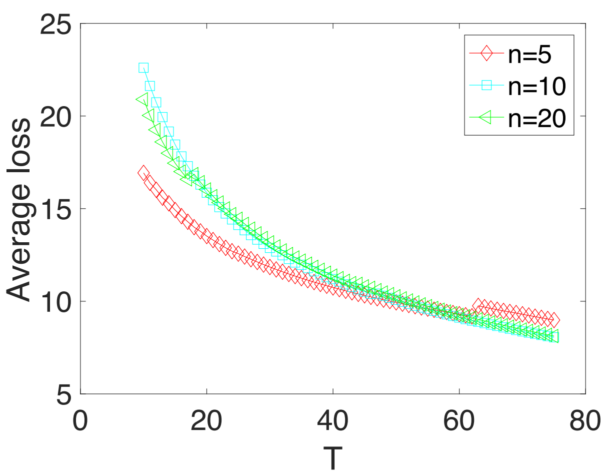
Second, figure 3 summarizes the effect of the network size on the performance of DOG. It shows that the performance of DOG is not sensitive to the network size, which confirms our theoretical result, that is, the average regret does not increase with the number of nodes.



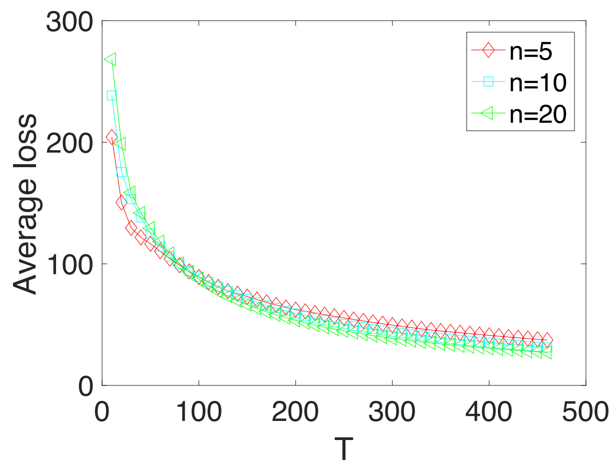
(a) *synthetic data*, ring topology



(b) *room-occupancy*, ring topology



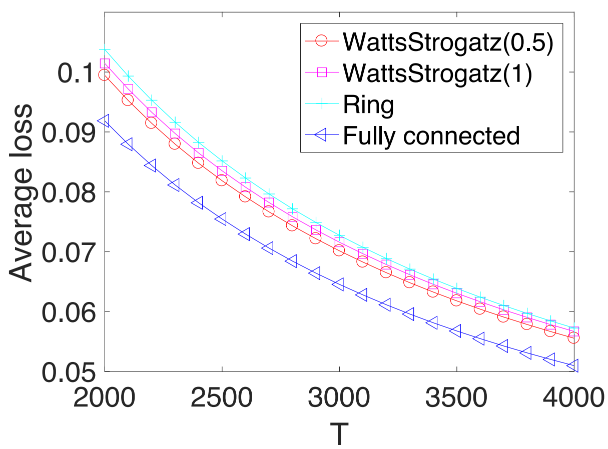
(c) *usenet2*, ring topology



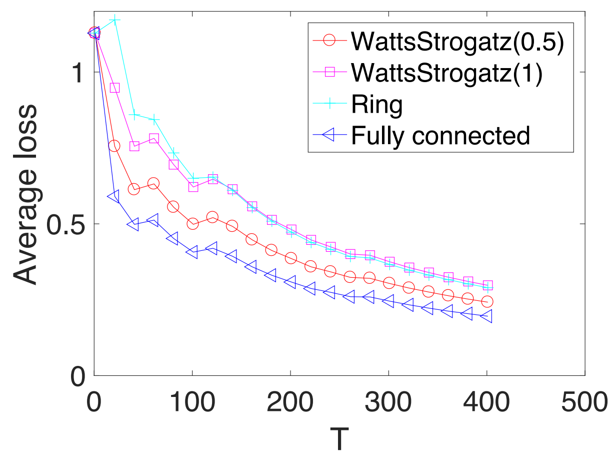
(d) *spam*, ring topology

Figure 3: The average loss yielded by DOG is insensitive to the network size.

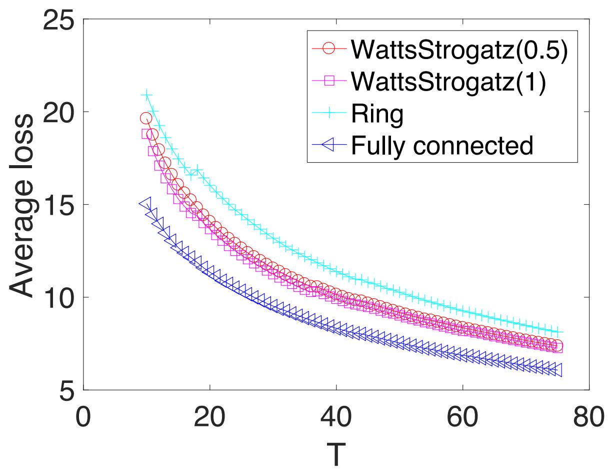
Third, figure 4 shows the effect of the topology of the network on the performance of DOG, for which four different topologies are used. Besides the ring topology, the *Fully connected* means all nodes are connected, where DOG de-generates to be COG. The topology *WattsStrogatz* represents a Watts-Strogatz small-world graph, for which we can use a parameter to control the number of random edges (set as 0.5 and 1 in this paper). The result shows *Fully connected* enjoyes the best performance, because that $\rho = 0$ for it while $\rho > 0$ for other topologies.



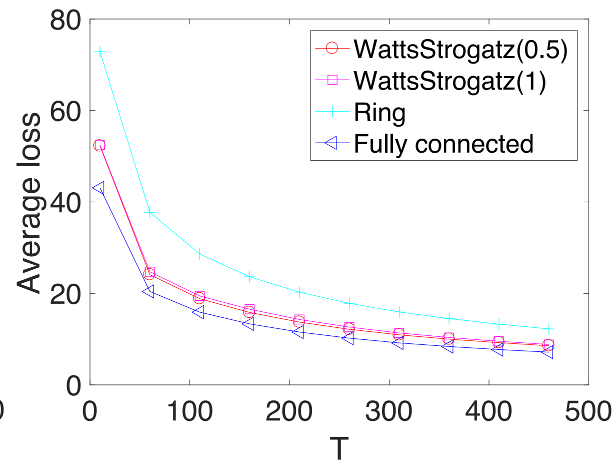
(a) *synthetic data*, 100 nodes



(b) *room-occupancy*, 20 nodes



(c) *usenet2*, 20 nodes



(d) *spam*, 20 nodes

Figure 4: The average loss yielded by DOG is insensitive to the topology of the network.

6 Conclusion

We investigate a new online learning problem in a decentralized network, where the loss incurs by both adversary and stochastic data. We provide a new analysis framework, which achieves sublinear regret. Extensive empirical studies varify the theoretical result.

References

D. Adamskiy, W. M. Koolen, A. Chernov, and V. Vovk. A closer look at adaptive regret. *Journal of Machine Learning Research*, 17(23):1–21, 2016.

M. Akbari, B. Gharesifard, and T. Linder. Distributed online convex optimization on time-varying directed graphs. *IEEE Transactions on Control of Network Systems*, 4(3):417–428, Sep. 2017.

A. S. Bedi, P. Sarma, and K. Rajawat. Tracking moving agents via inexact online gradient descent algorithm. *IEEE Journal of Selected Topics in Signal Processing*, 12(1):202–217, Feb 2018.

A. A. Benczúr, L. Kocsis, and R. Pálovics. Online Machine Learning in Big Data Streams. *CoRR*, 2018.

N. Cesa-Bianchi, P. Gaillard, G. Lugosi, and G. Stoltz. Mirror Descent Meets Fixed Share (and feels no regret). In *NIPS 2012*, page Paper 471, 2012.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research (JMLR)*, 12:2121–2159, 2011.

A. György and C. Szepesvári. Shifting regret, mirror descent, and matrices. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 2943–2951. JMLR.org, 2016.

A. György, T. Linder, and G. Lugosi. Tracking the Best of Many Experts. *Proceedings of Conference on Learning Theory (COLT)*, 2005.

A. Gyorgy, T. Linder, and G. Lugosi. Efficient tracking of large classes of experts. *IEEE Transactions on Information Theory*, 58(11):6709–6725, Nov 2012.

E. C. Hall and R. Willett. Dynamical Models and tracking regret in online convex programming. In *Proceedings of International Conference on International Conference on Machine Learning (ICML)*, 2013.

E. C. Hall and R. M. Willett. Online Convex Optimization in Dynamic Environments. *IEEE Journal of Selected Topics in Signal Processing*, 9(4):647–662, 2015.

E. Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3-4): 157–325, 2016.

M. Herbster and M. K. Warmuth. Tracking the best expert. *Machine Learning*, 32(2):151–178, Aug 1998.

A. Jadbabaie, A. Rakhlin, S. Shahrampour, and K. Sridharan. Online Optimization : Competing with Dynamic Comparators. In *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 398–406, 2015.

K.-S. Jun, F. Orabona, S. Wright, and R. Willett. Improved strongly adaptive online learning using coin betting. In A. Singh and J. Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 54, pages 943–951, 20–22 Apr 2017.

M. Kamp, M. Boley, D. Keren, A. Schuster, and I. Sharfman. Communication-efficient distributed online prediction by dynamic model synchronization. In *Proceedings of the 2014th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part I*, ECMLPKDD'14, pages 623–639, Berlin, Heidelberg, 2014. Springer-Verlag.

I. Katakis, G. Tsoumakas, and I. Vlahavas. Tracking recurring contexts using ensemble classifiers: An application to email filtering. *Knowledge and Information Systems*, 22(3):371–391, 2010.

A. Koppel, S. Paternain, C. Richard, and A. Ribeiro. Decentralized online learning with kernels. *IEEE Transactions on Signal Processing*, 66(12):3240–3255, June 2018.

S. Lee, A. Ribeiro, and M. M. Zavlanos. Distributed continuous-time online optimization using saddle-point methods. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 4314–4319, Dec 2016.

S. Lee, A. Nedi, and M. Raginsky. Coordinate dual averaging for decentralized online optimization with nonseparable global objectives. *IEEE Transactions on Control of Network Systems*, 5(1):34–44, March 2018.

M. Mohri and S. Yang. Competing with automata-based expert sequences. In A. Storkey and F. Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84, pages 1732–1740, 09–11 Apr 2018.

A. Mokhtari, S. Shahrampour, A. Jadbabaie, and A. Ribeiro. Online optimization in dynamic environments: Improved regret rates for strongly convex problems. In *Proceedings of IEEE Conference on Decision and Control (CDC)*, pages 7195–7201. IEEE, 2016.

J. Mourtada and O.-A. Maillard. Efficient tracking of a growing number of experts. *arXiv.org*, Aug. 2017.

A. Nedi, S. Lee, and M. Raginsky. Decentralized online optimization with global objectives and local communication. In *2015 American Control Conference (ACC)*, pages 4497–4503, July 2015.

S. Shahrampour and A. Jadbabaie. Distributed online optimization in dynamic environments using mirror descent. *IEEE Transactions on Automatic Control*, 63(3):714–725, March 2018.

S. Shalev-Shwartz. Online Learning and Online Convex Optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

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

C.-Y. Wei, Y.-T. Hong, and C.-J. Lu. Tracking the best expert in non-stationary stochastic environments. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Proceedings of Advances in Neural Information Processing Systems*, pages 3972–3980, 2016.

H.-F. Xu, Q. Ling, and A. Ribeiro. Online learning over a decentralized network through admm. *Journal of the Operations Research Society of China*, 3(4):537–562, Dec 2015.

F. Yan, S. Sundaram, S. V. N. Vishwanathan, and Y. Qi. Distributed autonomous online learning: Regrets and intrinsic privacy-preserving properties. *IEEE Transactions on Knowledge and Data Engineering*, 25 (11):2483–2493, Nov 2013.

T. Yang, L. Zhang, R. Jin, and J. Yi. Tracking Slowly Moving Clairvoyant - Optimal Dynamic Regret of Online Learning with True and Noisy Gradient. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2016.

C. Zhang, P. Zhao, S. Hao, Y. C. Soh, B. S. Lee, C. Miao, and S. C. H. Hoi. Distributed multi-task classification: a decentralized online learning approach. *Machine Learning*, 107(4):727–747, Apr 2018a.

L. Zhang, T. Yang, J. Yi, R. Jin, and Z.-H. Zhou. Improved Dynamic Regret for Non-degenerate Functions. In *Proceedings of Neural Information Processing Systems (NIPS)*, 2017a.

L. Zhang, T. Yang, rong jin, and Z.-H. Zhou. Dynamic regret of strongly adaptive methods. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 5882–5891, 10–15 Jul 2018b.

W. Zhang, P. Zhao, W. Zhu, S. C. H. Hoi, and T. Zhang. Projection-free distributed online learning in networks. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, pages 4054–4062, International Convention Centre, Sydney, Australia, 06–11 Aug 2017b.

Y. Zhao, S. Qiu, and J. Liu. Proximal Online Gradient is Optimum for Dynamic Regret. *CoRR*, cs.LG, 2018.

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 928–935, 2003.

Appendix

Proof to Theorem 1:

Proof.

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

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

① holds due to $\langle \mathbf{a},\mathbf{b}\rangle \le \frac{\eta}{2}\|\mathbf{a}\|^2 + \frac{1}{2\eta}\|\mathbf{b}\|^2$ holds for any $\eta>0$.

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

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

For $J_1(t)$, we have

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

① holds due to H_t has L-Lipschitz gradients, and $\bar{\mathbf{x}}_t = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_{i,t}$.

For $J_2(t)$, we have

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

① holds due to

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

② holds due to H_t has L Lipschitz gradients.

Therefore, we obtain

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

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

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

Therefore, we have

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

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

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

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

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

According to Lemma 5, we have

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

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

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

It is trivial to verify that (5) satisfies the first-order optimality condition of the optimization problem: $\min_{\mathbf{z}\in\mathbb{R}^d}\phi(\mathbf{z})$, that is,

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

We thus have

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

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

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

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

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

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

Besides, we have

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

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

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

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

Here, M the budget of the dynamics, which is defined in (2).

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

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

$$
\begin{aligned}
&\overset{②}{\leq}\eta T\left(n\beta G+(1-\beta)\sigma^2\right)+4n(1-\beta)\left(\mathop{\mathbb{E}}_{\Xi_{n,T}\sim\mathcal{D}_T}\sum_{t=1}^{T}\left(H_t(\bar{\mathbf{x}}_t)-H_t(\bar{\mathbf{x}}_{t+1})\right)\right)\\
&\quad+(1-\beta)\left(\frac{\beta}{2\eta}+L+2\eta L^2+4(1-\beta)^2L^2\eta\right)\frac{nT\eta^2G}{(1-\rho)^2}\\
&\quad+4n(1-\beta)\left(4T\beta^2\eta G+\frac{TGL\eta^2}{2}\right)+\frac{n}{2\eta}\left(4\sqrt{R}M+R\right)\\
&\leq\eta T\left(n\beta G+(1-\beta)\sigma^2\right)+4n(1-\beta)\left(\mathop{\mathbb{E}}_{\Xi_{n,T}\sim\mathcal{D}_T}\sum_{t=1}^{T}\left(H_t(\bar{\mathbf{x}}_t)-H_t(\bar{\mathbf{x}}_{t+1})\right)\right)\\
&\quad+\left(\frac{\beta}{2\eta}+L+2\eta L^2+4(1-\beta)^2L^2\eta\right)\frac{nT\eta^2G}{(1-\rho)^2}\\
&\quad+4n\left(4T\beta^2\eta G+\frac{TGL\eta^2}{2}\right)+\frac{n}{2\eta}\left(4\sqrt{R}M+R\right)\\
&\overset{③}{\leq}\eta T\left(n\beta G+(1-\beta)\sigma^2\right)+4n(1-\beta)T\eta G+\left(\frac{\beta}{2\eta}+L+2\eta L^2+4(1-\beta)^2L^2\eta\right)\frac{nT\eta^2G}{(1-\rho)^2}\\
&\quad+4n\left(4T\beta^2\eta G+\frac{TGL\eta^2}{2}\right)+\frac{n}{2\eta}\left(4\sqrt{R}M+R\right).
\end{aligned}
$$

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

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

② holds due to Lemma 1

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

③ holds due to

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

Re-arranging items, we have

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

$$
\begin{aligned}
&\leq \eta T n \beta \left(1 + \frac{1}{2(1-\rho)^2} + 16\beta\right) G + (1-\beta)\eta T \sigma^2 + 4n(1-\beta)T\eta G \\
&\quad + \left(\frac{L + 2\eta L^2 + 4(1-\beta)^2 L^2 \eta}{(1-\rho)^2} + 2L\right) n T \eta^2 G + \frac{n}{2\eta}\left(4\sqrt{R}M + R\right).
\end{aligned}
$$

It completes the proof.

□

Lemma 3. *Using Assumption 1, we have*

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

Proof.

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

It completes the proof. □

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

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

Proof.

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

Besides, we have

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

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

① holds due to

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

and

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

According to Lemma 3, we have

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

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

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

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

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

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

It completes the proof. □

Lemma 5. *Denote $\bar{\mathbf{x}}_t = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_{i,t}$. We have*

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

Proof. Denote

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

Recall that

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

Equivalently, we re-formulate the update rule as

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

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

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

Thus, we have

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

It completes the proof. □

Lemma 6 (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*

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

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

Lemma 7 (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

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

Proof of Lemma 1.

Proof. Recall that

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

and according to Lemma 5, we have

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

Denote

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

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

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

Therefore,

$$\begin{aligned}&\sum_{i=1}^{n} \|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\|^2\\ \overset{①}{=}& \sum_{i=1}^{n} \left\|\sum_{s=1}^{t-1} \eta\bar{\mathbf{G}}_s - \eta\mathbf{G}_s\mathbf{W}^{t-s-1}\mathbf{e}_i\right\|^2\end{aligned}$$

$$\overset{②}{=} \left\| \sum_{s=1}^{t-1} \eta \mathbf{G}_s \mathbf{v}_1 \mathbf{v}_1^{\mathrm{T}} - \eta \mathbf{G}_s \mathbf{W}^{t-s-1} \right\|_F^2$$
$$\overset{③}{\leq} \left(\eta \rho^{t-s-1} \left\| \sum_{s=1}^{t-1} \mathbf{G}_s \right\|_F \right)^2$$
$$\leq \left(\sum_{s=1}^{t-1} \eta \rho^{t-s-1} \left\| \mathbf{G}_s \right\|_F \right)^2 .$$

① holds due to $\mathbf{e}_i$ is a unit basis vector, whose i-th element is 1 and other elements are 0s. ② holds due to $\mathbf{v}_1 = \frac{\mathbf{1}_n}{\sqrt{n}}$. ③ holds due to Lemma 6.

Thus, we have

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

① holds due to Lemma 7. ② holds due to Lemma 3.

□

Proof of Lemma 2.

Proof.

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\xi_{i,t} \sim D_t} \left\| \nabla h_t(\mathbf{x}; \xi_{i,t}) \right\|^2 \\
= & \mathop{\mathbb{E}}_{\xi_{i,t} \sim D_t} \left\| \nabla h_t(\mathbf{x}; \xi_{i,t}) - \mathop{\mathbb{E}}_{\xi_{i,t} \sim D_t} \nabla h_t(\mathbf{x}; \xi_{i,t}) + \mathop{\mathbb{E}}_{\xi_{i,t} \sim D_t} \nabla h_t(\mathbf{x}; \xi_{i,t}) \right\|^2 \\
= & \mathop{\mathbb{E}}_{\xi_{i,t} \sim D_t} \left\| \nabla h_t(\mathbf{x}; \xi_{i,t}) - \mathop{\mathbb{E}}_{\xi_{i,t} \sim D_t} \nabla h_t(\mathbf{x}; \xi_{i,t}) \right\|^2 + \left\| \mathop{\mathbb{E}}_{\xi_{i,t} \sim D_t} \nabla h_t(\mathbf{x}; \xi_{i,t}) \right\|^2 \\
& + 2 \mathop{\mathbb{E}}_{\xi_{i,t} \sim D_t} \left\langle \nabla h_t(\mathbf{x}; \xi_{i,t}) - \mathop{\mathbb{E}}_{\xi_{i,t} \sim D_t} \nabla h_t(\mathbf{x}; \xi_{i,t}), \mathop{\mathbb{E}}_{\xi_{i,t} \sim D_t} \nabla h_t(\mathbf{x}; \xi_{i,t}) \right\rangle \\
= & \mathop{\mathbb{E}}_{\xi_{i,t} \sim D_t} \left\| \nabla h_t(\mathbf{x}; \xi_{i,t}) - \mathop{\mathbb{E}}_{\xi_{i,t} \sim D_t} \nabla h_t(\mathbf{x}; \xi_{i,t}) \right\|^2 + \left\| \mathop{\mathbb{E}}_{\xi_{i,t} \sim D_t} \nabla h_t(\mathbf{x}; \xi_{i,t}) \right\|^2 \\
\leq & \mathop{\mathbb{E}}_{\xi_{i,t} \sim D_t} \left\| \nabla h_t(\mathbf{x}; \xi_{i,t}) - \mathop{\mathbb{E}}_{\xi_{i,t} \sim D_t} \nabla h_t(\mathbf{x}; \xi_{i,t}) \right\|^2 + \mathop{\mathbb{E}}_{\xi_{i,t} \sim D_t} \left\| \nabla h_t(\mathbf{x}; \xi_{i,t}) \right\|^2 \\
\leq & \sigma^2 + G.
\end{aligned}$$

It thus completes the proof.

□