Gossip Online Learning: Exchanging Local Models to Track Dynamics

January 11, 2019

Abstract

1 Introduction

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

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

2 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 $\mathcal{O}\left(\log T\right)$ 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., 2018, 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. 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. Comparing with the previous result, our method obtains $\mathcal{O}\left(\sqrt{nTM}\right)$ dynamic regret (defined in (1)) for a decentralized online gradient descent. 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 limits its novelity for a general online learning task. Additionally, many decentralized online optimization and learning methods are proposed, for example, decentralized online multi-task learning [Zhang et al., 2018], 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 in the dynamic environment. Besides, Benczúr et al. [2018] reviews online learning methods for big data streams. Yan et al. [2013] provides necessary and sufficient conditions to preserve privacy for decentralized online learning methods.

2.2 Dynamic regret

3 Notations

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

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

respectively. For math brevity, we use the notation $\Xi_{n,T} \sim \mathcal{D}_{n,T}$ to represent that $\xi_{i,t} \sim D_{i,t}$ holds for any $i \in [n]$ and $t \in [T]$. $\mathbb{E}$ represents mathematical expectation. ∂ and ∇ represent sub-gradient and gradient operators, respectively. $\|\cdot\|$ represents the ℓ_2 norm in default.

4 Problem formulation

4.1 Setup

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

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

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

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

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

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

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

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

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

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

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

and

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

Recall that the previous definition of the dynamic regret is (1). Using (1), the classic online learning in a decentralized network only considers the loss function, i.e., $g_{i,t}$, incurred by the learning model on every node. Comparing with it, our definition of the dynamic regret, i.e., (2), still considers the loss function, i.e., H_t. It is incurred by a global model, which is used to let the decision variables, e.g., $\mathbf{x}_{i,t}$, have some good property in practical scenarios. We present some application scenarios to explain it in Section 4.2.

4.2 Application scenarios

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

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

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

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

Online learning with privacy protection. Suppose we want to conduct online learning on a decentralized network. But, there is a hacker who can sniff at the network, and obtains the transmitted data packages. To protect the privacy, we use a randomization encryption method to protect the local model before transmitting it in the network. In the case, $g_{i,t}(\mathbf{x}_{i,t})$ represents the loss incurred by the learning model. $h_t(\mathbf{x}_{i,t};\xi_{i,t})$ represents the loss incurred by some a randomization encryption method, e.g., objective perturbation [Chaudhuri et al., 2011, Wang et al., 2017], to protect the privacy. A small β means a strong guarantee for the data privacy.

Similarly, suppose we want to conduct online classification by using logistic regression model. Given an instance $\mathbf{a}_{i,t} \in \mathbb{R}^d$ and its label $\mathbf{y}_{i,t} \in \{1,-1\}$. In the case, $g_{i,t}(\mathbf{x}) = \log\left(1+\exp(-\mathbf{y}_{i,t}\mathbf{a}_{i,t}^{\mathrm{T}}\mathbf{x})\right)$. We use the objective perturbation strategy [Chaudhuri et al., 2011, Wang et al., 2017] to protect the privacy. Specifically, we let $h_t(\mathbf{x};\xi_{i,t}) = \mathbf{x}^{\mathrm{T}}\xi_{i,t}$, where $\xi_{i,t}$ is random noise, whose density is

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

Here, λ is a normalizing constant, $\delta_{i,t}$ is a known function of the constant $\epsilon_{i,t}$ for $\epsilon_{i,t}$-differential privacy [Dwork and Roth, 2014]. For example, when $\delta_{i,t} = \epsilon_{i,t}$, $\lambda =$. In the case, $H_t(\mathbf{x})$ makes the decision variable own privacy-preserving property.

Online recommendation with unreliable features. Suppose we want to decide whether to recommend music to Bob by using a public dataset consisting of historical browser records on Youtube. But, some values of features in those records are not reliable. For example, Alice's browser record is in the public dataset. But Alice does not want to let others know her real birthday and age. She submits random numbers for such information when signing up as an Youtube user. Note that those unreliable values, e.g., Alice's age and birthday, usually do not change, which is modeled by an unknown distribution. But, other reliable values, e.g., Alice's perference to music, may change over time, which is a classic setting for an online learning problem. In the case, $g_{i,t}(\mathbf{x}_{i,t})$ represents the loss incurred by those reliable features in the learing model, e.g., perference to music. $h_t(\mathbf{x}_{i,t};\xi_{i,t})$ represents the loss incurred by those unreliable features in the learing model, e.g., age and birthday. A small β means significant attention on those unreliable features.

Suppose we still use logistic regression to decide whether to recommend music to Bob. Without loss of generality, features corresponding to those unreliable values are denoted by the beginning s features. Given a user's behavior record $\mathbf{a}_{i,t}$ and its label $\mathbf{y}_{i,t} \in \{1,-1\}$. In the case, $g_{i,t}(\mathbf{x}) = \log\left(1+\exp\left(-\mathbf{y}_{i,t}\mathbf{a}_{i,t}^{\mathrm{T}}\hat{\mathbf{I}}\mathbf{x}\right)\right)$,

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

where $\hat{\mathbf{I}}$ is yielded by letting the first s diagonal elements of an identity matrix be 0s. $\xi_{i,t} = \check{\mathbf{I}}\mathbf{a}_{i,t}\mathbf{y}_{i,t}^{\mathrm{T}}$, and $h_t(\mathbf{x};\xi_{i,t}) = \log\left(1+\exp\left(-\xi_{i,t}^{\mathrm{T}}\mathbf{x}\right)\right)$, where $\check{\mathbf{I}}$ is yielded by letting the last $(d-s)$ diagonal elements of an identity matrix be 0s. Here, $\xi_{i,t}$ is drawn form an unknown distribution, that is, $\xi_{i,t} \sim D_{i,t}$, and $D_{i,t}$ usually changes unsignificant over t, or does not change over t. In the case, $H_t(\mathbf{x})$ allows the decision variable to represent different models to treat the unreliable and reliable features.

5 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**

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 $f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t})$.

Update:

4: Query a sub-gradient $\partial f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t})$.

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

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

6 Theoretical analysis

Assumption 1. *We make the following assumptions.*

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

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

  and

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

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

- *For any $i \in [n]$ and $t \in [T]$, we assume the function $f_{i,t}$ is convex, but may be non-smooth. Furthermore, we assume the function H_t has L-Lipschitz gradients. In brief, $g_{i,t}$ may be non-convex, non-smooth. H_t is smooth, but may be non-convex. $f_{i,t}$ is convex, but may be non-smooth.*

6.1 Main results

Theorem 1. *Denote $\bar{\mathbf{x}}_t = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_{i,t}$, and constants C_0 and C_1 by*

$$C_0 := \frac{1}{\sqrt{\beta^2+\eta}} + 4;$$

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

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

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\Xi_{n,T}\sim\mathcal{D}_{n,T}} \sum_{t=1}^{T}\sum_{i=1}^{n} f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t}) - f_{i,t}(\mathbf{x}_t^*;\xi_{i,t}) \\
\le& \eta T\left(n\beta G + (1-\beta)\sigma^2\right) + n(1-\beta)C_0\left(\mathop{\mathbb{E}}_{\Xi_{n,T}\sim\mathcal{D}_{n,T}} \sum_{t=1}^{T}\left(H_t(\bar{\mathbf{x}}_t) - H_t(\bar{\mathbf{x}}_{t+1})\right)\right) \\
&+ (1-\beta)\frac{nT\eta^2 GC_1}{(1-\rho)^2} + n(1-\beta)C_0\left(4T\beta^2\eta G + \frac{TGL\eta^2}{2}\right) + \frac{n}{2\eta}\left(4\sqrt{R}M + R\right).
\end{aligned}$$

Corollary 1. *Recall that*

$$C_0 = \frac{1}{\sqrt{\beta^2+\eta}} + 4.$$

Using Assumption 1, and choosing

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

in Algorithm 1, we have

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

6.2 Connections with the previous results

7 Empirical studies

7.1 Experimental settings

We simulate a decentralized network consisting of 5 nodes. Those nodes are connected by using a ring topology. Besides, we conduct online logistic regression by using three time series datasets: *room-occupancy*[2], *online-retail*[3], *BeijingPM2.5*[4], and a spam email dataset with the concept drift [Katakis et al., 2010]: *spam*[5] in the decentralized network. The data distribution of those datasets may change over time in those practial scenarios, leading to the change of the optimal learning model during online learing. In those dynamic environment, the dynamic regret is practial and necessary.

- *room-occupancy*. It collects features of a room including temperature, humidity, light, and CO2 for every minute between 02/02/2015 and 02/10/2015. Label of an instance is whether the room is occupied. Our goal is to learn a classification model to make a decision whether the room is occupied by using those features.

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

[2] https://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+
[3] https://archive.ics.uci.edu/ml/datasets/Online+Retail
[4] https://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data
[5] http://mlkd.csd.auth.gr/concept_drift.html

- *BeijingPM2.5*. It collects some weather features, e.g., teperature and pressure, and the PM2.5 data of US Embassy in Beijing hourly between 01/01/2010 and 12/31/2014. When the PM 2.5 index is larger than 100, the air quality is *bad*, otherwise, *good*. We want to train a binary clssification model to make a decision whether the air quality is good accoring to features such as temperature and pressure.

- *spam*. Every instance in the dataset is an email, where the frequence of every word in the dictionary is collected. But, the distribution of words changes over time, which is denoted by *concept drift*[Katakis et al., 2010]. We want to learn a classification model to make a decision whether an email is a spam.

All values of a feature have been normalized to be zero mean and one variance. The budget of dynamics, namely M is set to be 10. For the t-th iteration, the learning rate, namely η in Algorithm 1 is set to be $\sqrt{\frac{5M}{100t}}$. As we have shown in Section 4.2, we test the dynamic regret in those three application scenarios.

7.2 Communication efficient online logistic regression

The hyper-parameter to control the communication efficiency, namely λ_t is set to be a constant 0.1.

7.3 Online logistic regression with privacy protection

7.4 Online logistic regression with unrealiable features

References

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

K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. Differentially Private Empirical Risk Minimization. *Journal of Machine Learning Research*, 2011.

C. Dwork and A. Roth. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

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

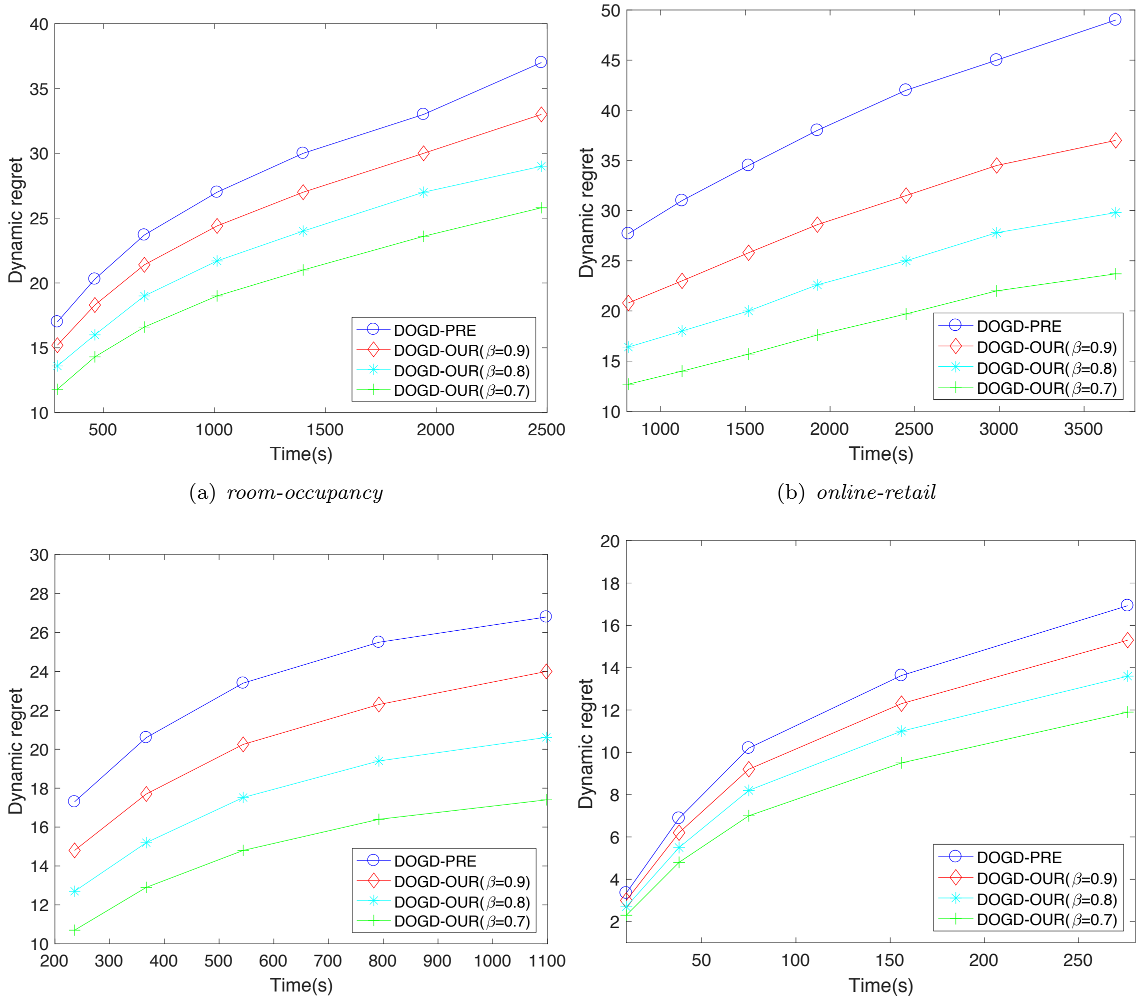
M. Kamp, M. Boley, D. Keren, A. Schuster, and I. Sharfman. Communication-efficient distributed online prediction by dynamic model synchronization. In *Proceedings of the 2014th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part I*, ECMLPKDD'14, pages 623–639, Berlin,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



(a) *room-occupancy*

(b) *online-retail*

(c) *BeijingPM2.5*

(d) *spam*

Figure 1: Comparsion of the dynamic regret by using communication efficient logistic regression in the decentralized network.

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.

D. Wang, M. Ye, and J. Xu. Differentially private empirical risk minimization revisited: Faster and more general. In *Advances in Neural Information Processing Systems 30*, pages 2722–2731. 2017.

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.

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 2018.

Appendix

Proof to Theorem 1:

Proof.

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

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

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

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

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

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

For $J_1(t)$, we have

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

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

For $J_2(t)$, we have

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

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

① holds due to

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

② holds due to H_t has L Lipschitz gradients.

Therefore, we obtain

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

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

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

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

Therefore, we have

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

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

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

According to Lemma 3, we have

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

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

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

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

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

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

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

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

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

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

Besides, we have

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

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

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

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

Here, M the budget of the dynamics, which is defined in (3).

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

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

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

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

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

② holds due to Lemma 4

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

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

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

It completes the proof.

□

Lemma 1. *Using Assumption 1, we have*

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

Proof.

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

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

It completes the proof. ∎

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

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

Proof.

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

Besides, we have

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

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

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

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

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

① holds due to

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

and

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

According to Lemma 1, we have

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

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

$$\mathop{\mathbb{E}}_{\Xi_{n,T}\sim\mathcal{D}_{n,T}} \sum_{t=1}^T H_t(\bar{\mathbf{x}}_{t+1})$$

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

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

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

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

$$\frac{\eta}{2}\mathop{\mathbb{E}}_{\Xi_{n,T-1}\sim\mathcal{D}_{n,T-1}} \sum_{t=1}^T \|\nabla H_t(\bar{\mathbf{x}}_t)\|^2 \tag{12}$$

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

It completes the proof. $\square$

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

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

Proof. Denote

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

Recall that

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

Equivalently, we re-formulate the update rule as

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

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

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

Thus, we have

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

It completes the proof. □

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

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

Proof. Recall that

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

and according to Lemma 3, we have

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

Denote

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

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

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

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

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

Therefore,

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

① holds due to $\mathbf{e}_i$ is a unit basis vector, whose i-th element is 1 and other elements are 0s. ② holds due to $\mathbf{v}_1 = \frac{\mathbf{1}_n}{\sqrt{n}}$. ③ holds due to Lemma 5.

Thus, we have

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

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

□

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

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

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

Lemma 6 (Appeared in Lemma 6 in [Tang et al., 2018]). *Given two non-negative sequences $\{a_t\}_{t=1}^{\infty}$ and $\{b_t\}_{t=1}^{\infty}$ that satisfying*

$$a_t = \sum_{s=1}^{t} \rho^{t-s} b_s,$$

with $\rho \in [0, 1)$, we have

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