Decentralized Online Learning: Take Benefits from Others' Data without Sharing Your Own and Track Global Dynamics

Anonymous Authors[1]

Abstract

Decentralized Online Learning (Online learning in decentralized networks) attracts more and more attention, since it is believed that Decentralized Online Learning can help the data providers cooperatively better solve their online problems without sharing their private data to a third party or other providers. Typically, the cooperation is achieved by letting the data providers exchange their models between neighbors, e.g., recommendation model. However, the best regret bound for a decentralized online learning algorithm is $\mathcal{O}\left(n\sqrt{T}\right)$, where n is the number of nodes (or users) and T is the number of iterations. This is clearly insignificant since this bound can be achieved *without* any communication in the networks. This reminds us to ask a fundamental question: *Can people really get benefit from the decentralized online learning by exchanging information?* In this paper, we studied when and why the communication can help the decentralized online learning to reduce the regret. Specifically, each loss function is characterized by two components: the adversarial component and the random component. Under this characterization, we show that decentralized online gradient (DOG) enjoys a regret bound $\mathcal{O}\left(n\sqrt{T}G + \sqrt{nT}\sigma\right)$, where G measures the magnitude of the adversarial component in the private data (or equivalently the local loss function) and σ measures the randomness within the private data. This regret suggests that people can get benefits from the randomness in the private data by exchanging private information. Another important contribution of this paper is to consider the dynamic regret – a more practical regret to track users' interest dynamics. Empirical studies are also conducted to validate our analysis.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

Decentralized Online Learning (or, online learning in decentralized networks) receives extensive attentions in recent years (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; Nedić et al., 2015; Lee et al., 2018; Benczúr et al., 2018; Yan et al., 2013). It assumes that computational nodes in a network can communicate between neighbors to minimize an overall cumulative regret. Each computational node, which could be a user in practice, will receive a stream of online losses that are usually determined by a sequence of examples that arrive sequentially. Formally, we can denote $f_{i,t}$ as the loss received by the i-th computational node among the networks at the t-th iteration. The goal of decentralized online learning usually is to minimize its static regret, which is defined as the difference between the cumulative loss (the sum of all the online loss over all the nodes and steps) suffered by the learning algorithm and that of the best model which can observe all the loss functions beforehand.

Decentralized online learning attracts more and more attentions recently, mainly because it is believed by the community that it enjoys the following advantages for real-world large-scale applications:

- (**Utilize all computational resource**) It can utilize the computational resource (of edging devices) by avoiding collecting all the loss functions (or equivalently data) to one central node and put all computational burden on a single node.
- (**Protect data privacy**) It can help many data providers collaborate to better minimize their cumulative loss, while at the same time protecting the data privacy as much as possible.

However, the current theoretical study does not explain why people need to use decentralized online learning, since the currently best regret result for decentralized online learning ($\mathcal{O}\left(n\sqrt{T}\right)$) for convex loss functions (Hosseini et al.,

2013; Yan et al., 2013)) is equal to the overall regret if each node (user) only runs local online gradient without any communication with others[1].

Therefore, this reminds us to ask a fundamental question:

Can people really get benefit with respect to the regret from the decentralized online learning by exchanging information?

In this paper, we mainly study when can the communication really help decentralized online learning to minimize its regret. Specifically, we distinguish two components in the loss function $f_{i,t}$: the adversary component and the random compoent. Then we prove that decentralized online gradient can achieve a static regret bound of $\mathcal{O}\left(n\sqrt{T}+\sqrt{nT}\sigma\right)$ (σ measures the randomness of the private data), where the first component of the bound is due to the adversary loss while the second component is due to the stochastic loss. Moreover, if a dynamic sequence of models with a budget M is used as the reference points, the dynamic regret of the decentralized online gradient is $\mathcal{O}\left(n\sqrt{TM}+\sqrt{nTM}\sigma\right)$. To Peilin: say something to motivate why we want to use dynamic regret. This shows the communication can help to minimize the stochastic losses, rather than the adversary losses. This result is further verified empirically by extensive experiments on several real datasets.

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 $D_{i,t}$, 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

$$\begin{aligned}\Xi_{n,T} =&\{\xi_{i,t}\}_{1\leq i\leq n, 1\leq t\leq T},\\ \mathcal{D}_{n,T} =&\{D_{i,t}\}_{1\leq t\leq T},\end{aligned} \tag{1}$$

  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.

- 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, $\mathbf{W}\mathbf{1} = \mathbf{1}$, and $\mathbf{1}^{\mathrm{T}}\mathbf{W} = \mathbf{1}^{\mathrm{T}}$. We use $\{\lambda_i\}_{i=1}^n$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ to represent its eigenvalues. Note that $\lambda_1 = 1$.

[1] n is the number of nodes or users and T is the total number of iterations. The regret of an online algorithm is $\mathcal{O}\left(\sqrt{T}\right)$ for convex loss functions (Hazan, 2016; Shalev-Swartz, 2012). Therefore, the overall regret is $n\sqrt{T}$ if all users do not communicate.

- ∇ represents gradient operator. $\|\cdot\|$ represents the ℓ_2 norm in default.
- $\lesssim$ represents "less than equal up to a constant factor".
- $\mathcal{A}$ represents the set of all online algorithms.
- $\mathbf{1}$ and $\mathbf{0}$ represent all the elements of a vector is 1 and 0, respectively.

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-Swartz, 2012; Bubeck, 2011). 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; Nedić 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, 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 (**??**)), 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 5). 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 novelty 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 gradient descent (Akbari et al., 2017), decentralized continuous-time online saddle-point method (Lee et al., 2016), decentralized online Nesterov's primal-dual method (Nedić et al.,

2015; Lee et al., 2018), and online distributed dual averaging(Hosseini et al., 2013). 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. Dynamic regret

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). For any online algorithm $A \in \mathcal{A}$, Zinkevich (2003) first define the dynamic regret by $\sum_{t=1}^{T}\left(f_t(\mathbf{x}_t) - f_t(\mathbf{x}_t^*)\right)$ subject to $\sum_{t=1}^{T-1}\left\|\mathbf{x}_{t+1}^* - \mathbf{x}_t^*\right\| \leq M$. They then propose an online gradient descent method, which yields $\mathcal{O}\left(\sqrt{TM} + \sqrt{T}\right)$ regret 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[2] for any given reference points. Besides, Zhao et al. (2018) presents that the lower bound of the dynamic regret is $\Omega\left(\sqrt{TM}\right)$. The previous definition of the regret is a special case of our new definition. Our analysis achieves the tight regret $\mathcal{O}\left(\sqrt{TM}\right)$ for the special case of $n = 1$ and $\sigma = 0$.

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.

[2] 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.

3. Problem formulation

Suppose that there are n users. Each user maintains a local predictive model, and only talk to his/her neighbors. Let $\mathbf{x}_{i,t}$ denote the local model for user i at iteration t. In iteration t user i applies the local model $\mathbf{x}_{i,t}$ to a function $f_{i,t}(\cdot;\xi_{i,t})$ and receives the loss $f_{i,t}(\cdot;\xi_{i,t})$. $\xi_{i,t}$'s are independent to each other in terms of i and t, charactering the *stochastic* component in the function $f_{i,t}(\cdot;\xi_{i,t})$, while the subscripts i and t of f indicate the *adversarial* component, for example, the user's profile, location, local time, and etc. **From Yawei:** The stochastic component in the function is usually caused by the potential realtion among local models. For example, users' perference to music may be impacted by a popular trend in the Internet at the same time.

Communication network. Users do not want to share the information to others, and can only share their private models to their neighbors (or friends). The communication network can be denoted by an undirected graph $\mathcal{G} = (\text{nodes:}[n], \text{edges:}E)$. Every node in $\mathcal{G}$ represents a user. For any node i with $i \in [n]$, its neighbour set contains all the directly adjacent nodes in $\mathcal{G}$.

Dynamic regret. For any online algorithm $A \in \mathcal{A}$, the commonly used regret used in online learning is *static*:

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

where $\mathbf{x}^*$ is a picked reference model.

It essentially assumes that the optimal model would not change over time. However, in many practical online learning application scenarios, the optimal model may evolve over time. For example, when we want to conduct music recommendation to a user, user's preference to music may change over time as his/her situation. Thus, the optimal model $\mathbf{x}^*$ should change over time. It leads to the dynamics of the optimal recommendation model. Therefore, for any online algorithm $A \in \mathcal{A}$, we choose to use the *dynamic* regret as the metric:

$$\mathcal{R}_T^A := \mathop{\mathbb{E}}_{\Xi_{n,T}\sim\mathcal{D}_{n,T}}\left[\sum_{i=1}^{n}\sum_{t=1}^{T} f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t})\right] - \mathop{\mathbb{E}}_{\Xi_{n,T}\sim\mathcal{D}_{n,T}}\min_{\{\mathbf{x}_t^\star\}_{t=1}^T\in\mathcal{L}_M^T}\left[\sum_{i=1}^{n}\sum_{t=1}^{T} f_{i,t}(\mathbf{x}_t^*;\xi_{i,t})\right], \tag{3}$$

where $\mathcal{L}_M^T$ is defined by

$$\mathcal{L}_M^T = \left\{ \{\mathbf{z}_t\}_{t=1}^T : \sum_{t=1}^{T-1} \|\mathbf{z}_{t+1} - \mathbf{z}_t\| \leq M \right\}.$$

$\mathcal{L}_M^T$ restricts how much the optimal model may change over time. When $M = 0$, the dynamic regret degenerates to the static regret.

4. Decentralized online gradient (DOG) algorithm

In the section, we introduce the DOG algorithm, followed by the analysis for the dynamic regret.

4.1. Algorithm description

Algorithm 1 DOG: Decentralized Online Gradient method.

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

1: Initialize $\mathbf{x}_{i,1} = \mathbf{0}$ for all $i \in [n]$.
2: **for** $t = 1, 2, ..., T$ **do**
3: // For all users (say the i-th node $i \in [n]$)
4: Query the neighbors' local models $\{\mathbf{x}_{j,t}\}_{j \in \text{user } i\text{'s neighbor set}}$.
5: Observe the loss function $f_{i,t}$, and suffer loss $f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t})$.
6: Query the gradient $\nabla f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t})$
7: Update the local model by

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

8: **end for**

In the DOG algorithm, users exchange their local models periodically. In each iteration, each user runs the following steps:

- (**Query**) Query the local models from his/her all neighbors;
- (**Gradient**) Apply the local model to $f_{i,t}(\cdot; \xi_{i,t})$ and obtain the gradient;
- (**Update**) Update the local model by taking average with neighbors' models followed by a gradient step.

The detailed description of the DOG algorithm can be found in Algorithm 1. $\mathbf{W} \in \mathbb{R}^{n \times n}$ is the confusion matrix of the graph $\mathcal{G} = (\text{nodes}{:}[n], \text{edges}{:}E)$. It is generated by the following steps.

1. For any node i with $i \in [n]$, if there is an edge $e_{ij} \in E$ between node i and one of its neighbour j, then $\mathbf{W}_{i,j} = 1$.
2. $\mathbf{W}$ is symmetric, that is, if $\mathbf{W}_{i,j} = 1$, then $\mathbf{W}_{j,i} = 1$. Every diagonal element of $\mathbf{W}$ is 1.
3. Normalize every row of $\mathbf{W}$, and make sure the sum of elements for every row is 1, that is $\mathbf{W}\mathbf{1} = \mathbf{1}$. Since $\mathbf{W}$ is symmetric, the sum of elements for every column is 1 as well, that is, $\mathbf{1}^{\mathrm{T}}\mathbf{W} = \mathbf{1}^{\mathrm{T}}$.

Denote

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

From the global view of point, the updating rule of DOG can be cast into the following form

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

Denote $\bar{\mathbf{x}}_t = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_{i,t}$. We can verify that $\bar{\mathbf{x}}_{t+1} = \bar{\mathbf{x}}_t - \frac{\eta}{n}\sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t}; \xi_{i,t})$ (see Lemma 2).

4.2. Dynamic regret of DOG

Next we show the dynamic regret of DOG in the following. Before that, we first make some common assumption used in our analysis.

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

Assumption 1. *We make following assumptions throughout this paper:*

- *For any $i \in [n]$, $t \in [T]$, and $\mathbf{x}$, there exist constants G and σ such that*

$$\mathop{\mathbb{E}}_{\xi_{i,t} \sim D_{i,t}} \|\nabla f_{i,t}(\mathbf{x}; \xi_{i,t})\|^2 \leq G^2,$$

and

$$\mathop{\mathbb{E}}_{\xi_{i,t} \sim D_{i,t}} \|\nabla f_{i,t}(\mathbf{x}; \xi_{i,t}) - \nabla F_{i,t}(\mathbf{x})\|^2 \leq \sigma^2.$$

- *For given vectors $\mathbf{x}$ and $\mathbf{y}$, we assume $\|\mathbf{x} - \mathbf{y}\|^2 \leq R$.*
- *For any $i \in [n]$ and $t \in [T]$, we assume the function $f_{i,t}$ is convex, and has L-Lipschitz gradient.*
- *The confusion matrix $\mathbf{W}$ is symmetric and doubly stochastic. Let ρ be $\rho := \max\{|\lambda_2(\mathbf{W})|, |\lambda_n(\mathbf{W})|\}$ and assume $\rho < 1$.*

G essentially gives the upper bound for the adversarial component in $f_{i,t}(\cdot;\xi_{i,t})$. The stochastic component is bounded by σ^2. Note that if there is no stochastic component, G is nothing but the upper bound of the gradient like the setting in many online learning literature. It is important for our analysis to split these two components, which will be clear very soon.

The last assumption about $\mathbf{W}$ is an essential assumption for the decentralized setting. The largest eigenvalue for a doubly stochastic matrix is 1. $1-\rho$ is the spectral gap, measuring how fast the information can propagate within the network (the larger the faster).

Now we are ready present the dynamic regret for DOG.

Theorem 1. *Let the constant C be*

$$C := \frac{L + 2\eta L^2 + 4L^2\eta}{(1-\rho)^2} + 2L.$$

Choosing $\eta > 0$ in Algorithm 1, under Assumption 1 we have

$$\begin{aligned}&\mathop{\mathbb{E}}_{\Xi_{n,T}\sim\mathcal{D}_{n,T}} \sum_{t=1}^{T}\sum_{i=1}^{n} f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t}) - f_{i,t}(\mathbf{x}_t^*;\xi_{i,t})\\ \leq& 20\eta T n G^2 + \eta T\sigma^2 + CnT\eta^2G^2 + \frac{n}{2\eta}\left(4\sqrt{R}M + R\right).\end{aligned}$$

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

Corollary 1. *Using Assumption 1, and choosing*

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

in Algorithm 1, we have

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

For simpler discussion, let us treat R, G, and $1-\rho$ as constants. The dynamic regret can simplified into $O(n\sqrt{MT}G + \sqrt{nMT}\sigma)$. If $M = 0$, the dynamic regret degenerates the static regret $O(n\sqrt{T}G + \sqrt{nT}\sigma)$. The discussion for the dynamic regret is conducted in the following

- (**Tightness.**) To see the tightness, we consider a few special cases:
  - ($\sigma = 0$ and $n = 1$.) It degenerates to the vanilla online learning setting but with dynamic regret. The implied static regret $O(\sqrt{TM})$ is consistent with the dynamic regret result in Zhao et al. (2018), which is proven to be optimum.
  - ($G = 0$ and $M = 0$.) It degenerates to the decentralized optimization scenario Duchi, John C et al. (2012); Tang et al. (2018). The static regret $O(\sqrt{nT}\sigma)$ implies the convergence rate $\sigma/\sqrt{nT}$, which is consistent with the result in Duchi, John C et al. (2012); Tang et al. (2018).
- (**Insight.**) Consider the baseline that all users do not communicate but only run local online gradient. It is not hard to verify that the static regret for this baseline approach is $O(n\sqrt{T}G + n\sqrt{T}\sigma)$. Comparing to the static regret ($O(n\sqrt{T}G + \sqrt{nT}\sigma)$) with iterative communication, the improvement is only on the stochastic component. Recall that G measures the magnitude of the adversarial component and σ measures the stochastic component. This result reveals an important observation that *the communication does not really help improve the adversarial component, only the stochastic component can benefit from the communication.* This observation makes quite sense, since if the users' private data are totally arbitrary, there is no reason they can benefit to each other by exchanging anything.
- (**Improve existing dynamic regret in decentralized setting.**) Shahrampour and Jadbabaie (2018) only considers the adversary loss, and provides $\mathcal{O}\left(n^{\frac{3}{2}}\sqrt{\frac{MT}{1-\rho}}\right)$ regret for DOG. 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.
- (**Improve existing static regret in the decentralized setting (Zhang et al., 2017b).**) When setting $M = 0$ in the regret defined in (2), Zhang et al. (2017b) provides $\mathcal{O}\left(nT^{\frac{3}{4}}\right)$ static regret for the decentralized online conditional gradient method. Our analysis shows that the regret can be improved to $\mathcal{O}\left(n\sqrt{T}G + \sqrt{nT}\sigma\right)$ by using DOG.

Next we discuss how close all local models $\mathbf{x}_{i,t}$'s close to their average at each time. The following result suggests that $\mathbf{x}_{i,t}$'s are getting closer and closer over iterations.

Theorem 2. *Recall $\bar{\mathbf{x}}_t = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_{i,t}$. Use Assumption 1, and choose*

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

in Algorithm 1. We have

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

The result suggests that $\mathbf{x}_{i,t}$ approaches to $\bar{\mathbf{x}}_t$ roughly in the rate $O(1/T)$, which is faster than the convergence of the averaged regret $O(1/\sqrt{T})$ from Corollary 1. For any online algorithm $A \in \mathcal{A}$, existing researeches (Zhang et al., 2017b) define the regret by using any local model, e.g., $\mathbf{x}_{j,t}$, instead of $\mathbf{x}_{i,t}$ on the i-th node. It is defined by

$$\widehat{\mathcal{R}}_T^A(\mathbf{x}_{j,t}) := \mathop{\mathbb{E}}_{\Xi_{n,T}\sim\mathcal{D}_{n,T}}\left[\sum_{i=1}^{n}\sum_{t=1}^{T} f_{i,t}(\mathbf{x}_{j,t};\xi_{i,t})\right] - \min_{\{\mathbf{x}_{i,t}^*\}_{t=1}^T\in\mathcal{L}_M^T} \mathop{\mathbb{E}}_{\Xi_{n,T}\sim\mathcal{D}_{n,T}}\left[\sum_{i=1}^{n}\sum_{t=1}^{T} f_{i,t}(\mathbf{x}_{i,t}^*;\xi_{i,t})\right],$$

where $\mathbf{x}_{j,t}$ is the local model for the j-th node with $j \in [n]$ at the time t. Inspired by Theorem 2, we find that the existing regret $\widehat{\mathcal{R}}_T^{\text{DOG}}(\mathbf{x}_{j,t})$ can be bounded by the following theorem.

Theorem 3. *Recall*

$$C := \frac{L + 2\eta L^2 + 4L^2\eta}{(1-\rho)^2} + 2L.$$

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

$$\begin{aligned}&\widehat{\mathcal{R}}_T^{\text{DOG}}(\mathbf{x}_{j,t})\\ \lesssim&\left(\frac{41Tn}{2} + \frac{nT}{2(1-\rho)^2}\right)\eta G^2 + \eta T\sigma^2\\ &+ CnT\eta^2G^2 + \frac{n}{2\eta}\left(4\sqrt{R}M + R\right)\end{aligned}$$

Choosing an appropriate learning rate η, we successfully obtain $\mathcal{O}\left(n\sqrt{T}G + \sqrt{nT}\sigma\right)$ regret for the regret $\widehat{\mathcal{R}}_T^{\text{DOG}}(\mathbf{x}_{j,t})$.

Corollary 2. *Using Assumption 1, and choosing*

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

in Algorithm 1, we have

$$\begin{aligned}&\widehat{\mathcal{R}}_T^{\text{DOG}}(\mathbf{x}_{j,t})\\ \lesssim& n\left(M + \sqrt{R}\right) + \frac{\sqrt{T\left(M + \sqrt{R}\right)(n^2G^2 + n\sigma^2)}}{1-\rho}.\end{aligned}$$

Using the definition of regret $\widehat{\mathcal{R}}_T^A(\mathbf{x}_{j,t})$ with $M = 0$, Zhang et al. (2017b) presents $\mathcal{O}\left(nT^{\frac{3}{4}}\right)$ static regret for decentralized online conditional gradient method, and Yan et al. (2013) presents $\mathcal{O}\left(n\sqrt{T}\right)$ static regret for decentralized autonomous online learning method. But, Theorem 3 and Corollary 2 show that our new analysis framework provides $\mathcal{O}\left(n\sqrt{TM}G + \sqrt{nTM}\sigma\right)$ dynamic regret for DOG, which improves the existing result in Zhang et al. (2017b), and extends both of those regret in the dynamic setting.

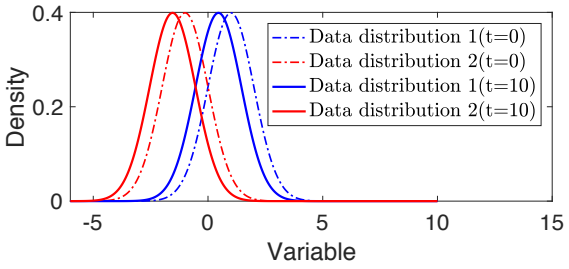


Figure 1. An illustration of the dynmaics caused by the time-varying distributions of data. Data distributions 1 and 2 satisify $N(1+\sin(t), 1)$ and $N(-1+\sin(t), 1)$, respectively. Suppose we want to conduct classification between data drawn from distributions 1 and 2, respectively. The optimal classification model should change over time.

5. Empirical studies

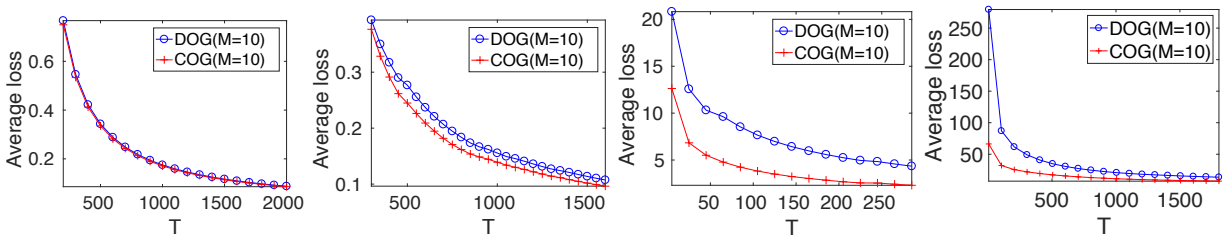
For simplicity, in the experiments we only consider online logistic regression with squared ℓ_2 norm regularization, i.e., $f_{i,t}(\mathbf{x};\xi_{i,t}) = \log\left(1+\exp(-\mathbf{y}_{i,t}\mathbf{A}_{i,t}^{\mathrm{T}}\mathbf{x})\right) + \frac{\gamma}{2}\|\mathbf{x}\|^2$, where $\gamma = 10^{-3}$ is a given hyper-parameter. Under this setting, we compare the proposed Decentralized Online Gradient method (DOG) and the Centralized Online Gradient method (COG).

M is fixed as 10 to determine the space of reference points. The learning rate η is tuned to be optimal for each dataset sperately. We evaluate the learning performance by measuring the average loss $\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t})$, instead of the dynamic regret $\mathbb{E}_{\Xi_{n,T}\sim\mathcal{D}_{n,T}}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t}) - f_{i,t}(\mathbf{x}_t^*)\right)$, since the optimal reference point $\{\mathbf{x}_t^*\}_{t=1}^T$ is the same for both DOG and COG .

5.1. Datasets

To test the proposed algorithm, we utilized a toy dataset and three real-world datasets, whose details are presented 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 $\tilde{\mathbf{A}}_j$ 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



(a) *synthetic data*, 10000 nodes,stochastic topology (b) *room-occupancy*, 5 nodes, ring topology (c) *usenet2*, 5 nodes, ring topology (d) *spam*, 5 nodes, ring topology

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

$\mathbf{y}_{i,t} = 1$, $\hat{\mathbf{A}}_{i,t}$ 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*[3], *usenet2*[4], and *spam*[5]. *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, DOG yields comparable performance with COG. 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 10000 nodes, where every node is stochasticly connected with other 15 nodes. For the three real datasets, 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, the performance of DOG is not sensitive to the network size, but sensitive to the variance of the stochastic data. Figure 3 summarizes the effect of the network size on the performance of DOG. We change the number of nodes from 5000 to 10000 on the synthetic dataset, and from 5 to 20 on the real datasets. The synthetic dataset is tested by using the stochastic topology, and those real datasets are tested by using the ring topology. Figure 3 draws the curves of average loss over time steps. We observe that the average loss curves are mostly overlapped with different nodes. It shows that DOG is robust to the network size (or number of users), which validates our theory, that is, the average regret does not increase with the number of nodes. Furthermore, we observe that the average loss becomes large with the increase of the variance of stochastic data, which validates our theoretical result nicely.

ρ	NC	FC	Ring	WS(1)	Ws(0.5)
synthetic data	1	0	0.99	0.37	0.58
real data	1	0	0.96	0.83	0.76

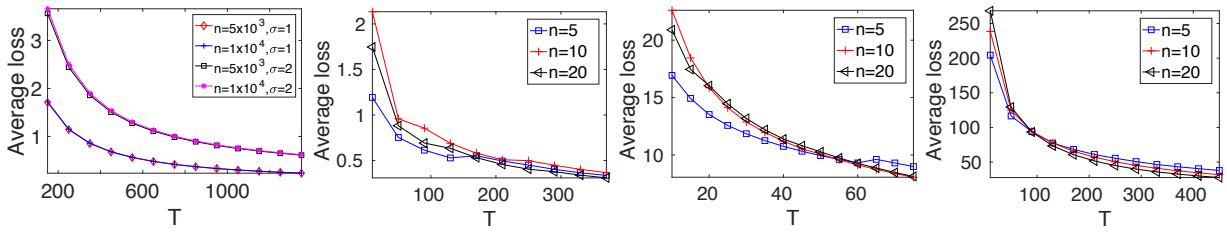
Table 1. ρ in different topologies used in our experiment. A large $1 - \rho$ represents good connectivity of the communication network. "NC" represents the *No connected* topology, "FC" represents the *Fully connected* topology, and "WS" represents the *WattsStrogatz* topology.

Third, the performance DOG is improved in a well-connected network. Figure 4 shows the effect of the topology of the network on the performance of DOG, where five different topologies are used. Besides the ring topology, the *No connected* topology means there are no edges in the network, and every node does not share its local model to others. The *Fully connected* topology 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 stochastic edges (set as 0.5 and 1 in this paper). The result shows *Fully connected* enjoys the best performance, because that $\rho = 0$ for it while $\rho > 0$ for other topologies. Specifically, ρ in those topologies is presented in Table 1. A small ρ leads to a good performance of DOG, which validates our theoretical result nicely.

[3] https://archive.ics.uci.edu/ml/datasets/Occupancy+Detection+

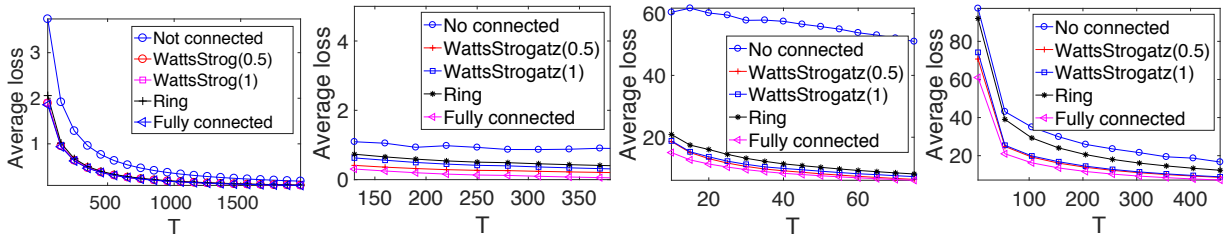
[4] http://mlkd.csd.auth.gr/concept_drift.html

[5] http://mlkd.csd.auth.gr/concept_drift.html



(a) *synthetic data*, stochastic 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.



(a) *synthetic data*, 10000 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 the online learning problem in a decentralized network, where the loss is incurred by both adversary and stochastic components. We define a new dynamic regret, and propose a decentralized online gradient method. By using the new analysis framework, the decentralized online gradient method achieves $\mathcal{O}\left(n\sqrt{TM}+\sqrt{nT}\sigma\right)$ regret. It shows that the communication is only effective to decrease the regret caused by the stochastic loss. Extensive empirical studies validates the theoretical results.

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.

S. Bubeck. Introduction to online optimization, December 2011.

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.

Duchi, John C, Agarwal, Alekh, and Wainwright, Martin J. Dual Averaging for Distributed Optimization - Convergence Analysis and Network Scaling. *IEEE Trans. Automat. Contr.*, 57(3):592–606, 2012.

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.

S. Hosseini, A. Chapman, and M. Mesbahi. Online distributed optimization via dual averaging. In *52nd IEEE Conference on Decision and Control*, pages 1484–1489, Dec 2013.

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. Nedić, 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. Nedić, 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. From the regret definition, we have

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

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

$$
I_1(t) = \left(\underbrace{\mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \frac{1}{n}\sum_{i=1}^{n} \langle \nabla f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t}), \mathbf{x}_{i,t} - \bar{\mathbf{x}}_t \rangle}_{J_1(t)} + \underbrace{\mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \left\langle \frac{1}{n}\sum_{i=1}^{n} \nabla f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t}), \bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1} \right\rangle}_{J_2(t)} \right).
$$

For $J_1(t)$, we have

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

① holds due to $F_{i,t}$ 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

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

$$
\begin{aligned}
&\leq \frac{\eta}{2} \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \left\| \frac{1}{n}\sum_{i=1}^{n} \left(\nabla f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t}) - \nabla F_{i,t}(\mathbf{x}_{i,t}) + \nabla F_{i,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 \\
&\leq \eta \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \left\| \frac{1}{n}\sum_{i=1}^{n} \left(\nabla f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t}) - \nabla F_{i,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 F_{i,t}(\mathbf{x}_{i,t})\right\|^2 \\
&\quad + \frac{1}{2\eta} \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \left\|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\right\|^2 \\
&\overset{①}{\leq} \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 F_{i,t}(\mathbf{x}_{i,t}) - \nabla F_{i,t}(\bar{\mathbf{x}}_t) + \nabla F_{i,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 \\
&\leq \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 F_{i,t}(\mathbf{x}_{i,t}) - \nabla F_{i,t}(\bar{\mathbf{x}}_t)\right)\right\|^2 \\
&\quad + 2\eta \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{t-1}} \left\|\nabla F_{i,t}(\bar{\mathbf{x}}_t)\right\|^2 + \frac{1}{2\eta} \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \left\|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\right\|^2 \\
&\leq \frac{\eta}{n}\sigma^2 + \frac{2\eta}{n} \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{t-1}} \sum_{i=1}^{n} \left\|\nabla F_{i,t}(\mathbf{x}_{i,t}) - \nabla F_{i,t}(\bar{\mathbf{x}}_t)\right\|^2 \\
&\quad + 2\eta \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{t-1}} \left\|\nabla F_{i,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 \\
&\overset{②}{\leq} \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 F_{i,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}
$$

① holds due to

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

② holds due to $F_{i,t}$ has L Lipschitz gradients.

Therefore, we obtain

$$
\begin{aligned}
&I_1(t) \\
=&(J_1(t) + J_2(t)) \\
=&\left(\frac{L}{n} \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{t-1}} \sum_{i=1}^{n} \left\|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\right\|^2 + \frac{\eta}{n}\sigma^2 + \frac{2\eta L^2}{n} \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{t-1}} \sum_{i=1}^{n} \left\|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\right\|^2 \right) \\
&+ \left(2\eta \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{t-1}} \left\|\nabla F_{i,t}(\bar{\mathbf{x}}_t)\right\|^2 + \frac{1}{2\eta} \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \left\|\bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1}\right\|^2 \right) \\
\leq&\left(\frac{L}{n} + \frac{2\eta L^2}{n}\right) \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{t-1}} \sum_{i=1}^{n} \left\|\mathbf{x}_{i,t} - \bar{\mathbf{x}}_t\right\|^2 + 2\eta \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{t-1}} \left\|\nabla F_{i,t}(\bar{\mathbf{x}}_t)\right\|^2
\end{aligned}
$$

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

Therefore, we have

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

Now, we begin to bound $I_2(t)$. Recall that the update rule is

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

According to Lemma 2, we have

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

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

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

It is trivial to verify that (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

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

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

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

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

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

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

$$\begin{aligned}
I_2(t) =& \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \left\langle \frac{1}{n}\sum_{i=1}^{n} \nabla f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t}), \bar{\mathbf{x}}_{t+1} - \mathbf{x}_t^* \right\rangle \\
\leq& \frac{1}{2\eta} \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \left(\lim_{\tau\to 0^+} \tau \left\| (\mathbf{x}_t^* - \bar{\mathbf{x}}_{t+1}) \right\|^2 + 2\left\langle \mathbf{x}_t^* - \bar{\mathbf{x}}_{t+1}, \bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_t \right\rangle \right) \\
=& \frac{1}{\eta} \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \left\langle \mathbf{x}_t^* - \bar{\mathbf{x}}_{t+1}, \bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_t \right\rangle \\
=& \frac{1}{2\eta} \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \left(\left\| \mathbf{x}_t^* - \bar{\mathbf{x}}_t \right\|^2 - \left\| \mathbf{x}_t^* - \bar{\mathbf{x}}_{t+1} \right\|^2 - \left\| \bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1} \right\|^2 \right). \qquad (6)
\end{aligned}$$

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 $I_2(t)$ over $t \in [T]$, we have

$$\begin{aligned}
\sum_{t=1}^{T} I_2(t) \leq& \frac{1}{2\eta} \mathop{\mathbb{E}}_{\Xi_{n,T}\sim\mathcal{D}_{n,T}} \left(4\sqrt{R} \sum_{t=1}^{T} \left\| \mathbf{x}_{t+1}^* - \mathbf{x}_t^* \right\| + \left\| \bar{\mathbf{x}}_1^* - \bar{\mathbf{x}}_1 \right\|^2 - \left\| \bar{\mathbf{x}}_T^* - \bar{\mathbf{x}}_{T+1} \right\|^2 \right) - \frac{1}{2\eta} \mathop{\mathbb{E}}_{\Xi_{n,T}\sim\mathcal{D}_{n,T}} \sum_{t=1}^{T} \left\| \bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1} \right\|^2 \\
\leq& \frac{1}{2\eta} \left(4\sqrt{R}M + R \right) - \frac{1}{2\eta} \mathop{\mathbb{E}}_{\Xi_{n,T}\sim\mathcal{D}_{n,T}} \sum_{t=1}^{T} \left\| \bar{\mathbf{x}}_t - \bar{\mathbf{x}}_{t+1} \right\|^2.
\end{aligned}$$

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

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

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

$$\overset{③}{\leq} \eta T\sigma^2 + 4nT\eta G^2 + \left(L + 2\eta L^2 + 4L^2\eta\right)\frac{nT\eta^2G^2}{(1-\rho)^2} + 4n\left(4T\eta G^2 + \frac{TG^2L\eta^2}{2}\right) + \frac{n}{2\eta}\left(4\sqrt{R}M + R\right).$$

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

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

② holds due to Lemma 5

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

③ holds due to

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

Re-arranging items, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\Xi_{n,T}\sim\mathcal{D}_{n,T}}\sum_{t=1}^{T}\sum_{i=1}^{n}f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t}) - f_{i,t}(\mathbf{x}_t^*;\xi_{i,t}) \\
\leq &20\eta TnG^2 + \eta T\sigma^2 + \left(\frac{L + 2\eta L^2 + 4L^2\eta}{(1-\rho)^2} + 2L\right)nT\eta^2G^2 + \frac{n}{2\eta}\left(4\sqrt{R}M + R\right).
\end{aligned}$$

It completes the proof.

□

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

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

Proof. We have

$$\mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}}F_{i,t}(\bar{\mathbf{x}}_{t+1})$$

$$
\begin{aligned}
&\leq \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{t-1}} F_{i,t}(\bar{\mathbf{x}}_t) + \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \langle \nabla F_{i,t}(\bar{\mathbf{x}}_t), \bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_t \rangle + \frac{L}{2} \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \|\bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_t\|^2 \\
&= \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{t-1}} F_{i,t}(\bar{\mathbf{x}}_t) + \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \left\langle \nabla F_{i,t}(\bar{\mathbf{x}}_t), -\frac{\eta}{n}\sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t}) \right\rangle + \frac{L}{2} \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \left\| \frac{\eta}{n}\sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t}) \right\|^2 \\
&= \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{t-1}} F_{i,t}(\bar{\mathbf{x}}_t) + \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{t-1}} \left\langle \nabla F_{i,t}(\bar{\mathbf{x}}_t), -\frac{\eta}{n}\sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t}) \right\rangle + \frac{L}{2} \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \left\| \frac{\eta}{n}\sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t}) \right\|^2 .
\end{aligned} \tag{8}
$$

Besides, we have

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{t-1}} \left\langle \nabla F_{i,t}(\bar{\mathbf{x}}_t), -\frac{\eta}{n}\sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t}) \right\rangle \\
&= \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{t-1}} \frac{\eta}{2} \left(\left\| \nabla F_{i,t}(\bar{\mathbf{x}}_t) - \frac{1}{n}\sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t}) \right\|^2 - \|\nabla F_{i,t}(\bar{\mathbf{x}}_t)\|^2 - \left\| \frac{1}{n}\sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t}) \right\|^2 \right) \\
&\leq \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{t-1}} \frac{\eta}{2} \left(\left\| \nabla F_{i,t}(\bar{\mathbf{x}}_t) - \frac{1}{n}\sum_{i=1}^n (\nabla f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t}) + \nabla F_{i,t}(\mathbf{x}_{i,t})) \right\|^2 \right) - \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{t-1}} \frac{\eta}{2} \|\nabla F_{i,t}(\bar{\mathbf{x}}_t)\|^2 \\
&\leq \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{t-1}} \frac{\eta}{2} \left(2 \left\| \nabla F_{i,t}(\bar{\mathbf{x}}_t) - \frac{1}{n}\sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t}) \right\|^2 + 2 \left\| \nabla F_{i,t}(\bar{\mathbf{x}}_t) - \frac{1}{n}\sum_{i=1}^n \nabla F_{i,t}(\mathbf{x}_{i,t}) \right\|^2 \right) \\
&\quad - \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{t-1}} \frac{\eta}{2} \|\nabla F_{i,t}(\bar{\mathbf{x}}_t)\|^2 \\
&\leq \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{t-1}} \frac{\eta}{2} \left(2 \left\| \nabla F_{i,t}(\bar{\mathbf{x}}_t) - \frac{1}{n}\sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t}) \right\|^2 + \frac{2}{n}\sum_{i=1}^n \|\nabla F_{i,t}(\bar{\mathbf{x}}_t) - \nabla F_{i,t}(\mathbf{x}_{i,t})\|^2 \right) \\
&\quad - \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{t-1}} \frac{\eta}{2} \|\nabla F_{i,t}(\bar{\mathbf{x}}_t)\|^2 \\
&\leq \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{t-1}} \frac{\eta}{2} \left(2 \left\| \nabla F_{i,t}(\bar{\mathbf{x}}_t) - \frac{1}{n}\sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t}) \right\|^2 + \frac{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 F_{i,t}(\bar{\mathbf{x}}_t)\|^2 \\
&\leq \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{t-1}} \frac{\eta}{2} \left(4\|\nabla F_{i,t}(\bar{\mathbf{x}}_t)\|^2 + 4 \left\| \frac{1}{n}\sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t}) \right\|^2 + \frac{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 F_{i,t}(\bar{\mathbf{x}}_t)\|^2 \\
&\overset{①}{\leq} \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{t-1}} \frac{\eta}{2} \left(8G^2 + \frac{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 F_{i,t}(\bar{\mathbf{x}}_t)\|^2 .
\end{aligned} \tag{9}
$$

① holds due to

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

Recall that

$$
\mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \|\nabla f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t})\|^2 \leq G^2. \tag{10}
$$

Substituting (9) and (10) into (8), and telescoping $t \in [T]$, we obtain

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

$$
\begin{aligned}
&\leq \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{t-1}} F_{i,t}(\bar{\mathbf{x}}_t) + \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{t-1}} \left\langle \nabla F_{i,t}(\bar{\mathbf{x}}_t), -\frac{\eta}{n}\sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t}) \right\rangle + \frac{L}{2} \mathop{\mathbb{E}}_{\Xi_{n,t}\sim\mathcal{D}_{n,t}} \left\| \frac{\eta}{n}\sum_{i=1}^n \nabla f_{i,t}(\mathbf{x}_{i,t};\xi_{i,t}) \right\|^2 \\
&\leq \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{t-1}} F_{i,t}(\bar{\mathbf{x}}_t) + \left(\mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{t-1}} \frac{\eta}{2}\left(8G^2 + \frac{2L^2}{n}\sum_{i=1}^n \left\|\bar{\mathbf{x}}_t - \mathbf{x}_{i,t}\right\|^2 \right) - \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{t-1}} \frac{\eta}{2}\left\|\nabla F_{i,t}(\bar{\mathbf{x}}_t)\right\|^2 \right) + \frac{G^2 L\eta^2}{2} \\
&= \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{t-1}} F_{i,t}(\bar{\mathbf{x}}_t) + \left(4\eta G^2 + \frac{L^2\eta}{n} \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{t-1}} \sum_{i=1}^n \left\|\bar{\mathbf{x}}_t - \mathbf{x}_{i,t}\right\|^2 - \mathop{\mathbb{E}}_{\Xi_{n,t-1}\sim\mathcal{D}_{t-1}} \frac{\eta}{2}\left\|\nabla F_{i,t}(\bar{\mathbf{x}}_t)\right\|^2 \right) + \frac{G^2 L\eta^2}{2}.
\end{aligned}
$$

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

$$
\begin{aligned}
&\frac{\eta}{2} \mathop{\mathbb{E}}_{\Xi_{n,T-1}\sim\mathcal{D}_{n,T-1}} \sum_{t=1}^T \left\|\nabla F_{i,t}(\bar{\mathbf{x}}_t)\right\|^2 \qquad (11) \\
\leq &\mathop{\mathbb{E}}_{\Xi_{n,T}\sim\mathcal{D}_{n,T}} \sum_{t=1}^T \left(F_{i,t}(\bar{\mathbf{x}}_t) - F_{i,t}(\bar{\mathbf{x}}_{t+1})\right) + 4T\eta G^2 + \frac{L^2\eta}{n} \mathop{\mathbb{E}}_{\Xi_{n,T-1}\sim\mathcal{D}_{n,T-1}} \sum_{t=1}^T\sum_{i=1}^n \left\|\bar{\mathbf{x}}_t - \mathbf{x}_{i,t}\right\|^2 + \frac{TG^2L\eta^2}{2}.
\end{aligned}
$$

It completes the proof. □

Lemma 2. *Denote* $\bar{\mathbf{x}}_t = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_{i,t}$. *We have*

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

Proof. Denote

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

Recall that

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

Equivalently, we re-formulate the update rule as

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

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

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

Thus, we have

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

It completes the proof. □

Lemma 3 (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 4 (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

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

Shahrampour and Jadbabaie (2018) investigates the dynamic regret of DOG, and provide the following sublinear regret.

Theorem 4 (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 $\mathcal{R}_T^{\mathrm{DOG}}$ is bounded by $\mathcal{O}\left(n^{\frac{3}{2}}\sqrt{\frac{MT}{1-\rho}}\right)$.*

As illustrated in Theorem 4, Shahrampour and Jadbabaie (2018) has provided a $\mathcal{O}\left(n\sqrt{nTM}\right)$ regret for DOG. Comparing with the regret in Shahrampour and Jadbabaie (2018), our analysis improves the dependence on n, which benefits from the following better bound of difference between $\mathbf{x}_{i,t}$ and $\bar{\mathbf{x}}_t$.

Lemma 5. *Using Assumption 1, and setting $\eta > 0$ in Algorithm 1, we have*

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

Proof. Recall that

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

and according to Lemma 2, we have

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

Denote

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

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

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

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

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

Therefore, we obtain

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

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

Thus, we have

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

① holds due to Lemma 4. □

Proof to Theorem 2:

Proof. Setting $\eta=\sqrt{\frac{(1-\rho)\left(nM\sqrt{R}+nR\right)}{nTG^2+T\sigma^2}}$ into Lemma 5, we finally complete the proof. □

Lemma 6. *Assume* $\left\|\nabla h_t(\mathbf{x};\xi_{i,t})\right\|^2\leq G$. *It implies*

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

Proof. We have

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

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

It thus completes the proof. □