On the Decentralized Stochastic Gradient Descent with Markov Chain Sampling

Article *in* IEEE Transactions on Signal Processing · July 2023

3 authors:

Tao Sun
National University of Defense Technology
83 PUBLICATIONS **792** CITATIONS
SEE PROFILE

Dongsheng li
National lab for parallel and distributed processing
217 PUBLICATIONS **2,372** CITATIONS
SEE PROFILE

Bao Wang
Michigan State University
75 PUBLICATIONS **610** CITATIONS
SEE PROFILE

On the Decentralized Stochastic Gradient Descent with Markov Chain Sampling

Tao Sun, Dongsheng Li, and Bao Wang

***Abstract*—The decentralized stochastic gradient method emerges as a promising solution for solving large-scale machine learning problems. This paper studies the decentralized Markov chain gradient descent (DMGD), a variant of the decentralized stochastic gradient method, which draws random samples along the trajectory of a Markov chain. DMGD arises when obtaining independent samples is costly or impossible, excluding the use of the traditional stochastic gradient algorithms. Specifically, we consider the DMGD over a connected graph, where each node only communicates with its neighbors by sending and receiving the intermediate results. We establish both ergodic and nonergodic convergence rates of DMGD, which elucidate the critical dependencies on the topology of the graph that connects all nodes and the mixing time of the Markov chain. We further numerically verify the sample efficiency of DMGD.**

***Index Terms*—Markov chain sampling, Gradient descent, Decentralization, Distributed machine learning, Convergence.**

I. Introduction

Distributed machine learning is a promising solution to solve large-scale machine learning tasks [1], [2]. In this paper, we consider solving the optimization problems collaboratively using m agents connected by an undirected graph. In particular, we focus on solving the following problem

$$\mathcal{F}(\mathbf{x}) := \min_{\mathbf{x}\in\mathbb{R}^n} \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{\xi(i)}\big(F(\mathbf{x};\xi(i))\big), \tag{1}$$

where $\mathbb{E}_{\xi(i)}\big(F(\mathbf{x};\xi(i))\big) := \int_{\Pi_i} F(\mathbf{x},\xi(i))d\Pi_i(\xi(i))$, and Ξ_i is a statistical sample space with probability distribution Π_i at node i[1], and $F(\cdot;\xi(i)) : \mathbb{R}^n \to \mathbb{R}$ is a closed function associated with $\xi(i) \in \Xi_i$. Problem (1) formulates various machine learning problems, including multi-agent machine learning. We focus on the cases where obtaining an independent and identically distributed (i.i.d.) sample $\xi(i)$ from Ξ_i is challenging or even impossible at every node i under a decentralized setting for solving the optimization problem in Equation (1). The problem under study arises naturally in machine learning, and we list two motivating examples in the next subsection.

T. Sun and D. Li are with College of Computer, National University of Defense Technology, Changsha, Hunan, China. Emails: nudtsuntao@163.com, dsli@nudt.edu.cn.

B. Wang is with Department of Mathematics, Scientific Computing and Imaging (SCI) Institute, University of Utah, Salt Lake City, Utah, USA. Email: wangbaonj@gmail.com.

D. Li is the corresponding authors.

Dongsheng Li and Tao Sun are supported in part by the National Science Foundation of China (62025208), Hunan Provincial Natural Science Foundation of China (2022JJ10065), and Young Elite Scientists Sponsorship Program by CAST (2022QNRC001).

[1]For the sake of presentation, we omit the underlying σ-algebra.

A. Motivating Examples

Example 1. Decentralized Pairwise Learning. Given a pairwise loss function $f(\mathbf{w};\mathbf{z},\hat{\mathbf{z}})$, where $\mathbf{z},\hat{\mathbf{z}} \sim \mathcal{D}$ are the data, and $\mathbf{w}$ denotes the parameters of the machine learning model to be learned. Then the following pairwise learning model $\min_{\mathbf{w}\in\mathbb{R}^n} f(\mathbf{w}) := \mathbb{E}_{\mathbf{z},\hat{\mathbf{z}}\sim\mathcal{D}} f(\mathbf{w};\mathbf{z},\hat{\mathbf{z}})$ describes various classical machine learning tasks, including metric learning [3], [4], [5], [6], AUC maximization [7], [8], [9], [10], and ranking problems [11], [12]. The online algorithm for pairwise learning model employ a "reuse" method [13] to update the parameters $\mathbf{w}$ as follows $\mathbf{w}^{k+1} = \mathbf{w}^k - \gamma\nabla f(\mathbf{w}^k;\mathbf{z}^k,\mathbf{z}^{k-1})$, where $(\mathbf{z}^k)_{k\geq 0}$ are the received data. The authors of [13] point out that $\{\xi^k := (\mathbf{z}^k,\mathbf{z}^{k-1})\}_k$ forms a Markov chain because ξ^k is only dependent on ξ^{k-1} but not on $(\xi^{k-2},\xi^{k-3},...,\xi^1)$. Then the decentralized online algorithm, with m nodes, for pairwise learning can be formulated as follows

$$\min_{\mathbf{w}\in\mathbb{R}^n} \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{[\mathbf{z}(i),\hat{\mathbf{z}}(i)]\sim\mathcal{D}_i} f(\mathbf{w};\mathbf{z}(i),\hat{\mathbf{z}}(i)),$$

which is a problem with the form in Problem (1).

Example 2. Decentralized Identification of Multi-linear Dynamics. Consider the following stochastic linear dynamical system $\mathbf{x}^{t+1} = \mathbf{A}\mathbf{x}^t + \xi^t$, $\mathbf{y}^{t+1} = \mathbf{B}\mathbf{y}^t + \varsigma^t$ where $\xi^t \sim \mathcal{D}_1, \varsigma^t \sim \mathcal{D}_2$ are i.i.d. samples drawn from a given distribution. It is evident that $(\mathbf{x}^t)_{t\geq 0}$ and $(\mathbf{y}^t)_{t\geq 0}$ are Markov chains. The papers [14], [15] consider the following minimization problem $\min_{\mathbf{W}} \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\hat{\mathcal{D}}}\|\mathbf{W}\mathbf{x}-\mathbf{y}\|^2$, where $\hat{\mathcal{D}}$ is the stationary distribution of $(\mathbf{x}^t,\mathbf{y}^t)_{t\geq 0}$. The online algorithm for solving the above minimization problem can be formulated as $\mathbf{W}^{k+1} = \mathbf{W}^k - \gamma(\mathbf{W}^k\mathbf{x}^k - \mathbf{y}^k)[\mathbf{x}^k]^\top$, which is indeed an SGD with Markov chain sampling.

Let us further consider the following stochastic dynamical system $\mathbf{x}^{t+1}(i) = \mathbf{A}(i)\mathbf{x}^t(i) + \xi^t(i)$, $\mathbf{y}^{t+1}(i) = \mathbf{B}(i)\mathbf{y}^t(i) + \varsigma^t(i)$ for $i = 1,2,\ldots,m$, where $\xi^t(i),\varsigma^t(i)$ are i.i.d. samples drawn from some distribution. We denote the stationary distribution of $(\mathbf{x}^t(i),\mathbf{y}^t(i))_{t\geq 0}$ as $\mathcal{D}_i$. Then we can formulate the online algorithm for identifying multi-linear dynamics as follows [14]

$$\min_{\mathbf{W}} \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{[\mathbf{x}(i),\mathbf{y}(i)]\sim\mathcal{D}_i}\|\mathbf{W}\mathbf{x}(i) - \mathbf{y}(i)\|^2.$$

Therefore, decentralized SGD with Markov chain sampling is a natural algorithm for solving the above optimization problem.

B. Previous Decentralized Algorithms Lack Sample Efficiency

The existing algorithm for solving Problem (1) is decentralized SGD (DSGD) with Markov chain sampling [16], which can be written as follows

$$\mathbf{x}^{k+1}(i) = \sum_{l\in\mathcal{N}(i)} w_{i,l}\mathbf{x}^k(l) - \gamma_k \nabla F(\mathbf{x}^k(i);\xi^k(i)), \qquad (2)$$

where $\mathbf{x}^k(i)$ denotes the parameters at node i in iteration k, and $w_{i,l}$ is the (i,l)-th entry of the mixing matrix $\mathbf{W}$ (see Definition 1), and $\nabla F(\mathbf{x}^k(i);\xi^k(i))$ is a *nearly unbiased* stochastic gradient of $\mathbb{E}_{\xi(i)}\big(F(\mathbf{x}^k(i);\xi(i))\big)$ obtained at node i. In the following, we explain how to implement DSGD under the Markov chain samplings setting. By using the Markov chain, to perform one iteration of Equation (2), each node i has to generate a sequence of samples $\xi^1(i), \xi^2(i), \ldots, \xi^T(i)$ and only uses the last one $\xi^k(i) := \xi^T(i)$ for updating the local parameters following Equation (2).

According to [17, Theorem 4.9], to get a sample that is nearly i.i.d., one needs to simulate the Markov chain for a sufficiently long time, i.e., a large T. For this reason, we call the iteration Equation (2) with $\xi^k(i) := \xi^T(i)$ as DSGD-T. Therefore, applying iteration Equation (2) to solve Problem (1) over the distributed nodes lacks sample efficiency. In particular, implementing DSGD-T for solving Problem (1) requires regenerating a Markov chain at each node per iteration, which can be computationally prohibitive. In contrast, DMGD can find the near-optimal solution of Problem (1) with sample efficiency.

C. Preliminaries

1) Notation: For a vector $\mathbf{x} \in \mathbb{R}^N$, we denote its local copy at node i as $\mathbf{x}(i)$. For a squared matrix $\mathbf{A}$, we denote its i-th eigenvalue as $\lambda_i(\mathbf{A})$. For a matrix $\mathbf{A} = (a_{i,j})_{N\times n}$, we denote its Frobenius norm and infinity norm respectively as $\|\mathbf{A}\|$ and $\|\mathbf{A}\|_\infty$, i.e., $\|\mathbf{A}\| := \sqrt{\sum_{i=1}^N \sum_{j=1}^n a_{i,j}^2}$[2] and $\|\mathbf{A}\|_\infty := \max_{i,j}|a_{i,j}|$. For a positive semidefinite matrix $\mathbf{B}$, we denote $\|\mathbf{B}^{\frac{1}{2}}\mathbf{A}\|$ as $\|\mathbf{A}\|_{\mathbf{B}}$. We define the σ-algebra as $\chi^k := \sigma(\mathbf{x}^0, \mathbf{x}^1, \ldots, \mathbf{x}^k)$. We use $\mathbb{E}[\cdot]$ to denote the expectation with respect to the underlying probability measure, i.e., $\mathbb{E}[\cdot] = \mathbb{E}[\cdot \mid \chi^k]$.

2) Discretization of Problem (1)*:* Suppose all distributions $(\Pi_i)_{1\le i\le m}$, in Problem (1), are supported on a set of M points, $\mathbf{y}^{1,i}, \ldots, \mathbf{y}^{M,i}$ (for Π_i)[3]. We define the functions as $f_i^j(\mathbf{x}) := M\cdot\mathrm{Prob}(\xi = \mathbf{y}^{i,j})\cdot F(\mathbf{x};\mathbf{y}^{i,j})$, and thus Equation (1) becomes the following finite-sum optimization problem

$$\text{minimize}_{\mathbf{x}\in\mathbb{R}^n} \quad f(\mathbf{x}) := \frac{1}{m}\sum_{i=1}^m f_i(\mathbf{x}), \qquad (3)$$

where $f_i(\mathbf{x}) := \frac{1}{M}\sum_{j=1}^M f_i^j(\mathbf{x})$ is the loss function of the i-th node.

Denote $(j_{i,k})_{k\ge 0} \subseteq \{1, 2, \ldots, M\}$ as the trajectory of the Markov chain in the i-th node and k-th iteration. We use a connected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertex set $\mathcal{V} = \{1, ..., m\}$ and edge set $\mathcal{E} \subseteq \mathcal{V}\times\mathcal{V}$. Any edge $(i,l) \in \mathcal{E}$ represents a communication channel between nodes i and l. Moreover, let

$$\mathbf{j}_k := \begin{bmatrix} j_{1,k} \\ j_{2,k} \\ \vdots \\ j_{m,k} \end{bmatrix}, \ \mathbf{x} := \begin{bmatrix} \mathbf{x}(1)^\top \\ \mathbf{x}(2)^\top \\ \vdots \\ \mathbf{x}(m)^\top \end{bmatrix}, \ \mathbf{x}^k := \begin{bmatrix} \mathbf{x}^k(1)^\top \\ \mathbf{x}^k(2)^\top \\ \vdots \\ \mathbf{x}^k(m)^\top \end{bmatrix},$$
$$\mathbf{u}^k := \left[\nabla f_1^{j_{1,k}}(\mathbf{x}^k(1)), \ldots, \nabla f_m^{j_{m,k}}(\mathbf{x}^k(m))\right]^\top, \qquad (4)$$

where $\mathbf{x}(i)$ is the variable in the i-th node, $\mathbf{x}^k(i)$ is the k-th iterate in the i-th node, and $\mathbf{u}^k$ is the collection of the stochastic gradients in the k-th iteration.

3) Mixing matrix: The mixing matrix is frequently used in decentralized optimization. In many cases, it can be designed by the users according to the given graph. Formally, it is defined as follows.

Definition 1: The mixing matrix $\mathbf{W} = [w_{ij}] \in \mathbb{R}^{m\times m}$ is assumed to have the following properties: (1) *(Graph)* If $i \ne j$ and $(i,j) \notin \mathcal{E}$, then $w_{ij} = 0$, otherwise, $w_{ij} > 0$; (2) *(Symmetry)* $\mathbf{W} = \mathbf{W}^\top$; (3) *(Null space property)* $\mathrm{null}\{\mathbf{I} - \mathbf{W}\} = \mathrm{span}\{\mathbf{1}\}$; (4) *(Spectral property)* $\mathbf{I} \succeq \mathbf{W} \succ -\mathbf{I}$.

Since $\mathbf{W}$ is a symmetric matrix, its eigenvalues are all real and can be sorted in non-increasing order. That is, we can sort the eigenvalues of $\mathbf{W}$ as $\lambda_1(\mathbf{W}) = 1 > \lambda_2(\mathbf{W}) \ge \cdots \ge \lambda_m(\mathbf{W}) > -1$. Also, we denote $\lambda(\mathbf{W}) := \max\{|\lambda_2(\mathbf{W})|, |\lambda_m(\mathbf{W})|\}$.

4) Markov chain: In the following, we recall several definitions, properties, and existing results of the finite-state time-homogeneous Markov chain, which will be used in the proposed algorithms.

Definition 2: A stochastic process $X_1, X_2, ...$ in a finite state space $\{1, 2, \ldots, n\}$ is called a time-homogeneous Markov chain with transition matrix $\mathbf{H} \in \mathbb{R}^{n\times n}$ if for $i, j \in \{1, 2, \ldots, n\}$ and $i_0, i_1, \ldots, i_{k-1} \in \{1, 2, \ldots, n\}$ with $k \in \mathbb{N}$, we have $\mathbb{P}(X_{k+1} = j \mid X_0 = i_0, \ldots, X_k = i) = \mathbb{P}(X_{k+1} = j \mid X_k = i) = \mathbf{H}_{i,j}$.

Denote the probability distribution of X_k as the non-negative row vector $\pi^k = (\pi_1^k, \pi_2^k, \ldots, \pi_n^k)$, i.e., $\mathbb{P}(X_k = j) = \pi_j^k$ and π satisfies $\sum_{i=1}^n \pi_i^k = 1$. For the time-homogeneous Markov chain, it holds that $\pi^k = \pi^{k-1}\mathbf{H}$ and $\pi^k = \pi^{k-1}\mathbf{H} = \cdots = \pi^0\mathbf{H}^k$ for $k \in \mathbb{N}$, where $\mathbf{H}^k$ denotes the k-th power of $\mathbf{H}$.

A Markov chain is irreducible if, for any $i, j \in \{1, 2, \ldots, n\}$, there exists k such that $(\mathbf{H}^k)_{i,j} > 0$. State $i \in \{1, 2, \ldots, n\}$ is said to have a period d if $\mathbf{H}_{i,i}^k = 0$ whenever k is *not* a multiple of d and d is the largest integer with this property. If $d = 1$, then we say state i is aperiodic. If every state is aperiodic, the Markov chain is said to be aperiodic. Any time-homogeneous, irreducible, and aperiodic Markov chain has a stationary distribution $\pi^* = \lim_k \pi^k = [\pi_1^*, \pi_2^*, \ldots, \pi_n^*]$ with $\sum_{i=1}^n \pi_i^* = 1$ and $\min_i\{\pi_i^*\} > 0$, and $\pi^* = \pi^*\mathbf{H}$. It also holds that

$$\lim_k \mathbf{H}^k = \left[(\pi^*)^\top, (\pi^*)^\top, \ldots, (\pi^*)^\top\right]^\top =: \Pi^* \in \mathbb{R}^{n\times n}. \quad (5)$$

The largest eigenvalue of $\mathbf{H}$ is 1, and the corresponding eigenvector is π^*.

Mixing time is an important notion of the Markov chain, which describes how long a Markov chain evolves until its

[2]When $N = 1$ or $n = 1$, $\|\cdot\|$ is then the L_2 norm of a vector.

[3]For the sake of exposition, we assume the same cardinal number of the support set for different distributions.

Algorithm 1 DSGD-T

Require: parameters $(\gamma_k)_{k\geq 0}$, the mixing matrix $\mathbf{W}$
Initialization: $(\mathbf{x}^0(i))_{1\leq i\leq m}$ are i.i.d. selected from the ball $\mathbf{B}(\mathbf{0}, B)$ centered at $\mathbf{0}$ with radius B
for $k = 1, 2, \ldots$
for $i = 1, 2, \ldots, m$
1. **Resample** a Markov chain $j_0^k(i), \ldots, j_T^k(i)$
2. Collect $\mathbf{x}^k(l)$ with $l \in \mathcal{N}(i)$
3. Update $\mathbf{x}^{k+1}(i)$ via

$$\mathbf{x}^{k+1}(i) = \sum_{l\in\mathcal{N}(i)} w_{i,l}\mathbf{x}^k(l) - \gamma_k \nabla f_i^{j_T^k(i)}(\mathbf{x}^k(i)).$$

end for
end for

Algorithm 2 Decentralized Markov Gradient Descent (DMGD)

Require: parameters $(\gamma_k)_{k\geq 0}$, the mixing matrix $\mathbf{W}$
Initialization: $(\mathbf{x}^0(i))_{1\leq i\leq m}$ are i.i.d. selected from the ball $\mathbf{B}(\mathbf{0}, B)$ centered at $\mathbf{0}$ with radius B
for $k = 1, 2, \ldots$
for $i = 1, 2, \ldots, m$
1. **Sample** $j_{i,k}$ via a Markov chain
2. Collect $\mathbf{x}^k(l)$ with $l \in \mathcal{N}(i)$
3. Update $\mathbf{x}^{k+1}(i)$ via

$$\mathbf{x}^{k+1}(i) = \sum_{l\in\mathcal{N}(i)} w_{i,l}\mathbf{x}^k(l) - \gamma_k \nabla f_i^{j_{i,k}}(\mathbf{x}^k(i)).$$

end for
end for

current state has a distribution that is close to its stationary distribution. The literature on various kinds of mixing times is mostly about reversible Markov chains, i.e., the Markov chains that satisfy $\pi_i \mathbf{H}_{i,j} = \pi_j \mathbf{H}_{j,i}$. With basic matrix analysis, the mixing time introduced by [18] provides a direct relationship between k and the deviation of the distribution of the current state from the stationary distribution; see Lemma 1 in the Appendix for details.

D. Our Proposed Algorithm: Decentralized Markov Gradient Descent

In this subsection, we present the discrete formulation of DMGD, i.e., the scheme for solving the finite-sum problem in Equation (3). In the k-th iteration of DMGD, the i-th node performs the local iteration as follows

$$\mathbf{x}^{k+1}(i) = \sum_{l\in\mathcal{N}(i)} w_{i,l}\mathbf{x}^k(l) - \gamma_k \nabla f_i^{j_{i,k}}(\mathbf{x}^k(i)). \tag{6}$$

In each iteration of DMGD, each node calculates the local gradient along the Markov chain trajectory $(j_{i,k})_{k\geq 0}$, and then communicates with its neighbors $\mathcal{N}(i)$ using a weighted average $\sum_{l\in\mathcal{N}(i)} w_{i,l}\mathbf{x}^k(l)$ to update the local parameters. Here, $w_{i,l}$ is the (i,l)-th element of the mixing matrix. It is easy to see that if the trajectory of the Markov chain satisfies the uniform sampling, Equation (6) then reduces to the DSGD. In the above iteration, the stepsize γ_k needs to go to zero to guarantee the algorithm's convergence.

Notice that in the discrete case, i.e., for solving the finite-sum optimization problem, DSGD-T can be written as follows

$$\mathbf{x}^{k+1}(i) = \sum_{l\in\mathcal{N}(i)} w_{i,l}\mathbf{x}^k(l) - \gamma_k \nabla f_i^{j_T^k(i)}(\mathbf{x}^k(i)), \tag{7}$$

where $[j_0^k(i), j_1^k(i), \ldots, j_T^k(i)]$ is the Markov chain resampled in node i for the kth iteration. As T is large enough, $j_T^k(i)$ is close to the uniform i.i.d. sampling on the set $\{1, 2, \ldots, M\}$. We provide the pseudocode of DSGD-T and DMGD in **Algorithm 1** and **Algorithm 2**, respectively.

E. Related Works

In this part, we briefly review three lines of related works: decentralized optimization, decentralized stochastic optimization, and Markov chain gradient descent.

1) Decentralized optimization: Decentralized algorithms have been originally studied in control and signal processing communities, e.g., calculating the mean of data distributed over multiple sensors [19], [20], [21], [22]. Decentralized (sub)gradient descent (DGD) algorithms for the finite-sum optimization $\min_{\mathbf{x}\in\mathbb{R}^d} \sum_{i=1}^m f_i(\mathbf{x})$ have been studied by [23], [24], [25], [26], [27]. Instead of directly solving the original problem, DGD applies the gradient descent to solve the surrogate problem $F(\mathbf{X}) + \frac{1}{2\alpha}\text{tr}(\mathbf{X}^\top(\mathbf{I}-\mathbf{W})\mathbf{X})$, where $\mathbf{X} := \left([\mathbf{x}_i]_{1\leq i\leq m}\right)^\top \in \mathbb{R}^{m\times d}$, $F(\mathbf{X}) := \sum_{i=1}^m f_i(\mathbf{x}_i)$, $\alpha > 0$ is the penalty parameter, and $\mathbf{W}$ is the mixing matrix associated with the graph. Thus, DGD converges to an inexact solution. To fix this, the dual information is leveraged in recent works such as decentralized ADMMs and primal-dual algorithms [28], [29], [30], [31]. Although DGD is slower than decentralized ADMMs and primal-dual algorithms in convex settings, the scheme of DGD is much simpler and, therefore, easier to extend to the nonconvex, online, and delay-tolerant settings [32], [33], [34], [35].

2) Decentralized stochastic optimization: Decentralized SGD (DSGD) has been studied recently as a generalization of the classical decentralized optimization for deterministic optimization problems. By assuming a local Poisson clock for each agent, asynchronous gossip algorithms is proposed by [36], in which each worker randomly selects part of its neighbors to communicate with. In fact, these algorithms use random communication graphs. Decentralized algorithms with random communication graphs for the constrained problem is introduced by [37], and the subgradient counterpart is given by [38]. In recent works of [39], [40], [16], the theoretical convergence complexity analysis of convex and nonconvex DSGD has been established. [39] present the complexity analysis for a stochastic decentralized algorithm. [40] design a stochastic decentralized algorithm by recruiting the dual information and providing the related computational complexity analysis. In the latter paper [16], the authors show the speedup when the number of nodes is increased. And in paper [41], the authors propose the asynchronous DSGD. The generalization analysis of DSGD is established by [42]. In [43], DSGD has been modified for federated learning. DSGD has been developed in different applications under different settings; until 2020, an elegant algorithmic framework with dynamical graph topology

and local updates have been proposed by [44]. There are several recent works on developing communication-efficient variants of DSGD [45], [46], [47], [48], [49].

3) Markov chain gradient descent: While i.i.d. samples are not always available in stochastic optimization, recent works have focused on analyzing stochastic algorithms following a single trajectory of the Markov chain or other general ergodic processes. The key challenge of analyzing Markov chain gradient descent (MCGD) is to deal with the biased expectation of gradients. The ergodic convergence results have been established by [50], [51]. Specifically, [50], [51] study the conditional expectation with a sufficiently large delay which is sufficiently close to the gradient. [52] prove the almost sure convergence under the diminishing stepsizes $\gamma_k = 1/k^q$, $2/3 < q \leq 1$. [53] improve the convergence results with larger stepsizes $\gamma_k = 1/\sqrt{k}$ in the sense of ergodic convergence. In all the works above, the Markov chain needs to be reversible, and the functions have to be convex. In [18], the non-ergodic convergence of MGD has been shown in the nonconvex case with non-reversible Markov chain, but the algorithm needs to be implemented in a centralized fashion. The boundedness assumption of stochastic gradient or the iterate in MCGD is removed by [54], [55]. The provably accelerated version of MCGD in the convex case is proposed by [56]. In [57], the authors propose the adaptive MCGD with theoretical convergence guarantees.

II. Contributions and Technical Challenges

A. Our Contributions

The primary contribution of this paper is the development of the DMGD accompanied by performance analysis. In contrast to the well-known DSGD, *DMGD leverages Markov chain sampling rather than uniform random sampling*, which gains sample efficiency. For the first-order DMGD, each node uses a Markov chain trajectory to sample a gradient and then communicates with its neighbors to update the local parameters.

We establish the non-ergodic convergence analysis of the DMGD and their ergodic convergence rates. The results show that the DMGD converges at the same rate as the centralized MCGD. Some novel results are developed based on new techniques and approaches developed in this paper. We use varying mixing time rather than fixed ones to get stronger results in general cases. The numerical results demonstrate that DMGD outperforms DSGD in terms of sample efficiency.

Although our proof requires the use of the delay expectation employed by [18], our analysis is substantially different from that of [18]. This is because the convergence analysis of DMGD is established on estimating the successive difference on the average of iterates in all nodes, which does not apply the objective function in this paper. The proof of DMGD is built on the average of all nodes' parameters, which is also quite different from MCGD. To this end, several techniques are developed to characterize the difference between the average and nodes' parameters under Markov chain sampling.

B. Key Challenges for Analyzing DMGD

The main challenge of this paper is to integrate decentralized SGD with Markov chain sampling. Markov chain sampling is neither cyclic nor i.i.d. stochastic. For any large K, it is still possible that a sample is never visited within some $k+1, ..., k+K$ iterations. For a fixed node i, unless the local graph $\mathcal{G}_i$ is complete, there are nodes l, h not connected by an edge. Hence, given $j_{i,k-1} = l_i$, it is *impossible* to have $j_{i,k} = h_i$. So, no matter how one selects the sampling probability and stepsize γ_k, we generally *do not* have $\mathbb{E}_{\mathbf{j_{k-1}}}(\gamma_k \mathbf{u}^k) = C(\sum_{i=1}^m \nabla f_i(\mathbf{x}^k(i)))$ for any constant C. This fact, unfortunately, breaks down all the existing analyses of stochastic decentralized optimization since all existing analyses need a non-vanishing probability such that each node can be sampled.

Moreover, the difficulty of analyzing DMGD differs from the centralized case because the convergence analysis of DMGD is established by estimating the successive difference on the average of iterates in all nodes, which does not apply the objective function in this paper. Consequently, decentralized methods use different mathematical characterizations of their convergence. In particular, the norm of gradients is employed to characterize the convergence of the centralized case, i.e., we analyze the convergence of $\min_{1\leq k\leq K} \mathbb{E}\|\nabla f(\mathbf{x}^k)\|^2$ for the centralized case. In contrast, we characterize the convergence of DMGD using the average of all local iterations, i.e., $\min_{1\leq k\leq K} \mathbb{E}\|\nabla f(\overline{\mathbf{x}^k})\|^2$ with $\overline{\mathbf{x}^k}$ being defined in Equation (8) in the following context. The iterative scheme of $\mathbf{x}^k$ and $\overline{\mathbf{x}^k}$ can be written as follows $\mathbf{x}^{k+1} = \mathbf{x}^k - \gamma_k \nabla f_{i_k}(\mathbf{x}^k)$, $\overline{\mathbf{x}^{k+1}} = \overline{\mathbf{x}^k} - \gamma_k \sum_{i=1}^m \nabla f_i^{j_{i,k}}(\mathbf{x}^k(i))/m$. Notice that the centralized scheme is directly applied to $\mathbf{x}^k$, while the decentralized one uses the local iterates $(\mathbf{x}^k(i))_{1\leq i\leq m}$. Another technical difference between this work and [18] is that we further consider the mini-batch version of DMGD; in contrast, [18] only consider updating using a single sample.

III. Convergence analysis of DMGD

In this section, we present the theoretical convergence results of DMGD with finite-state Markov chains, and our analysis builds on the following assumptions.

Assumption 1: The function f_i is bounded below; that is, $\min f_i > -\infty$, $\forall i \in \{1, 2, \ldots, m\}$.

Assumption 2: The gradient of f_i^j is uniformly bounded; that is, there exists a constant $B > 0$ such that $\|\nabla f_i^j(\mathbf{x})\| \leq B$, for $\forall i \in \{1, 2, \ldots, m\}$, $\forall j \in \{1, 2, \ldots, M\}$.

Assumption 3: The gradient of f_i^j is Lipschitz continuous with L_i^j, i.e., $\|\nabla f_i^j(\mathbf{x}) - \nabla f_i^j(\mathbf{y})\| \leq L_i^j \|\mathbf{x} - \mathbf{y}\|$, and $\forall i \in \{1, 2, \ldots, m\}$, $\forall j \in \{1, 2, \ldots, M\}$. Moreover, we denote $L := \max_{1\leq i\leq m, 1\leq j\leq M}\{L_i^j\}$.

Assumption 4: The Markov chains in all nodes are time-homogeneous, irreducible, and aperiodic, which have the same transition matrix $\mathbf{H}$ and the same stationary distribution.[4]

Following the routines in the stochastic decentralized optimization community, the convergence of the algorithm is described by the quantity below

$$\overline{\mathbf{x}^k} := \frac{1}{m}\sum_{i=1}^m \mathbf{x}^k(i). \tag{8}$$

[4]We require all nodes to employ the Markov chain with same transition matrix $\mathbf{H}$. This setting is for the convenience of presentations in the proofs and can be modified as different Markov chains for different nodes.

Theorem 1: Suppose Assumptions 1-4 hold, and the stepsizes are selected as follows

$$\gamma_k = \frac{1}{(k+1)^\theta}, \quad \frac{1}{2} < \theta < 1. \tag{9}$$

For $(\mathbf{x}^k)_{k\geq 0}$ generated by DMGD, we have the following nonergodic convergence result

$$\lim_k \mathbb{E}\|\nabla f(\overline{\mathbf{x}^k})\| = 0. \tag{10}$$

Moreover, the ergodic convergence rate is given by

$$\begin{aligned}
&\min_{1\leq k\leq K} \mathbb{E}\|\nabla f(\overline{\mathbf{x}^k})\|^2 \\
&= \frac{c_1(\mathbf{H}) + c_2(\mathbf{H})\sum_{t=\tau(\mathbf{H})}^{+\infty} \frac{3\ln(4MC_{\mathbf{H}}B^2t)\ln t\cdot\gamma_t^2}{\ln^2(1/\lambda(\mathbf{H}))} + f(\overline{\mathbf{x}^0}) - \min f}{K^{1-\theta}} \\
&= \mathcal{O}\Big(\frac{1 + \frac{1}{\ln^2(1/\lambda(\mathbf{H}))}\cdot[1 + \frac{1}{\sqrt{m}(1-\lambda(\mathbf{W}))}] + f(\overline{\mathbf{x}^0}) - \min f}{K^{1-\theta}}\Big),
\end{aligned} \tag{11}$$

where $\lambda(\mathbf{H}) := \frac{\max\{|\lambda_2(\mathbf{H})|, |\lambda_{\min}(\mathbf{H})|\}+1}{2} \in [0,1)$, $\lambda_2(\mathbf{H})$ and $\lambda_{\min}(\mathbf{H})$ denote the second and smallest eigenvalue of $\mathbf{H}$, respectively, $c_1(\mathbf{H}) := \sum_{t=1}^{\tau(\mathbf{H})-1}\gamma_t B^2 + B^3L[\tau(\mathbf{H})]^4/(4m) + BL\sum_{t=1}^{\infty}\gamma_t^2 + \sum_{t=1}^{\infty}\frac{\gamma_t}{2t}$ with $\tau(\mathbf{H})$ being a constant dependent on $\mathbf{H}$ is defined by Equation (37), $c_2(\mathbf{H}) := 2B/\sqrt{m} + B^3L/m + 2B^2/m + \frac{2B^2LC_{\mathbf{W}}}{\sqrt{m}}$, and $C_{\mathbf{W}}$ is a constant given by Equation (52).

Theorem 1 uses complicated forms for several constants because we want to derive the explicit formula for the upper bound of the convergence. Due to the bias of the Markov chain sampling, we use delay expectation techniques in the proofs. To get the explicit formulation of the upper bound, we need to determine how large the explicit delay is needed, resulting in complicated forms.

In Theorem 1, the functions are unnecessary to be convex. Indeed, it is more challenging to prove Equation (10) than Equation (11). The descent on a Lyapunov function and Schwarz's inequality imply Equation (10), while to prove Equation (11) requires a technical lemma, which was first given in [32] and generalized by [18]. An special case is that $m = 1$ and $\mathbf{W}$ is the identity matrix $\mathbf{I}$, then DMGD reduces to the classical MCGD. But Theorem 1 cannot cover the existing convergence results of MCGD. [18] estimate the convergence of MCGD with the following stepsize constraints

$$\sum_{k=1}^{+\infty}\gamma_k = +\infty, \quad \sum_{k=1}^{+\infty}\ln^2 k\cdot\gamma_k^2 < +\infty. \tag{12}$$

The stepsize Equation (9) can satisfy Equation (12) but not vice versa.

Theorem 1 provides a favorable nonergodic convergence result, i.e., Equation (10). The result indicates that $(\mathbb{E}\|\nabla f(\overline{\mathbf{x}^k})\|)_{k\geq 0}$ converges to 0, allowing us to directly use the K-th iterate as the output rather than the average or the minimizer of $(\|\nabla f(\overline{\mathbf{x}^k})\|)_{0\leq k\leq K}$. The ergodic convergence rate describes the speed of DMGD, dependent on m, $\overline{\mathbf{x}^0}$, $\lambda(\mathbf{H})$, and $\lambda(\mathbf{W})$. As the number of nodes m increases, the upper bound decreases. Such a phenomenon is similar to the minibatch SGD because DMGD uses m-minibatch data but with decentralized updating and Markov chains sampling. Inappropriate initializations (i.e., $f(\overline{\mathbf{x}^0}) - \min f$ is large) increase the bound and thus hurt the convergence, which coincides with our intuition. Notice that $\lambda(\mathbf{H})$ characterizes the speed of the Markov chain converges to stationary speed: a smaller $\lambda(\mathbf{H})$ indicates a faster speed. From the above bound, we can see a faster Markov chain also yields a faster DMGD. However, the Markov chain comes from the nature of data and is usually not controlled by users. The above theoretical results also show that the graph structure affects the convergence: a smaller $\lambda(\mathbf{W})$ means a better convergence rate. A natural question is what kind of graph would one expect for the optimal convergence? This question can be mathematically formulated as finding $\mathbf{W}$ with the smallest $\lambda(\mathbf{W})$ for a given graph $\mathcal{G}$. It is a very complicated optimization problem, which is not our focus for this paper, and related results can be found in [58].

As both $\lambda(\mathbf{H})$ and $\lambda(\mathbf{W})$ are closed to 1, the convergence rate is then dominated by $\mathcal{O}\Big(\frac{1}{\sqrt{m}\ln^2(1/\lambda(\mathbf{H}))(1-\lambda(\mathbf{W}))K^{1-\theta}}\Big)$. This bound can be improved if we use the following stepsize rule

$$\gamma_k = \frac{\sqrt{\ln(1/\lambda(\mathbf{H}))}}{(k+1)^\theta}, \quad \frac{1}{2} < \theta < 1. \tag{13}$$

Proposition 1: Let conditions of Theorem 1 hold and use the stepsizes in Equation (13), then we have

$$\begin{aligned}
&\min_{1\leq k\leq K} \mathbb{E}\|\nabla f(\overline{\mathbf{x}^k})\|^2 \\
&\quad = \mathcal{O}\left(\frac{\frac{1}{\sqrt{m}\ln(1/\lambda(\mathbf{H}))(1-\lambda(\mathbf{W}))} + \frac{1+f(\overline{\mathbf{x}^0})-\min f}{\sqrt{\ln(1/\lambda(\mathbf{H}))}}}{K^{1-\theta}}\right).
\end{aligned}$$

When $\lambda(\mathbf{W})$ is not close to 1, Proposition 1 indicates that to reach the ϵ-error, i.e., $\mathbb{E}\|\nabla f(\overline{\mathbf{x}^k})\|^2 \leq \epsilon$, the iteration complexity of DMGD is $K = \mathcal{O}\Big(\frac{[\ln(1/\lambda(\mathbf{H})]^{\frac{1}{2(1-\theta)}}}{\epsilon^{\frac{1}{1-\theta}}}\Big)$. Notice that the iteration complexity of DMGD is almost the same as that of MCGD, given by $\mathcal{O}\Big(\frac{\ln(1/\lambda(\mathbf{H})}{\epsilon^2}\Big)$, because θ can be arbitrarily close to $1/2$. Although the stepsize Equation (13) can reduce the upper bound of the iteration complexity when $\lambda(\mathbf{H})$ is close to 1, we usually cannot use this stepsize rule in practice since $\lambda(\mathbf{H})$ is usually unknown.

The convergence results in Theorem 1 are built on Assumption 4, i.e., the Markov chain is time-homogeneous, irreducible, and aperiodic. Assumption 4 gives the geometric mixing time property, which is crucial for the subsequent analysis. We stress that the results can be extended to other Markov chains. For example, [52] and [53] show the geometric mixing time for other Markov chains under extra assumptions.

IV. Analysis on Continuous State Space

In this part, we consider the case that $\Pi_1, \Pi_2, \ldots, \Pi_m$ are continuous probability distributions, i.e., we consider the problem in Equation (1). In this case, we consider time-homogeneous and reversible infinite-state Markov chains, and Theorem 4.9 in [17] indicates that the mixing time of the Markov chain enjoys a geometric decay. Mathematically, such a geometric decay can be written as

$$\|\delta^k\|_\infty \leq C\cdot\lambda^k, \quad \text{as} \quad k \geq 0, \tag{14}$$

where δ^k still denotes the deviation matrix $\Pi^* - \mathbf{H}^k$. Here C and λ are constants determined by the Markov chain.

Let $\xi^0(i), \xi^1(i), \ldots$ be the trajectory of the Markov chain of the i-th node and $(\mathbf{x}^0(i))_{1\le i\le m}$ are i.i.d. selected from the ball $\mathbf{B}(\mathbf{0}, B)$ centered at $\mathbf{0}$ with radius B. By defining

$$\mathbf{d}^k := \left[\nabla F(\mathbf{x}^k(1); \xi^k(1)); \ldots; \nabla F(\mathbf{x}^k(m); \xi^k(m))\right]^\top,$$

the global scheme is then of the following form

$$\mathbf{x}^{k+1} = \mathbf{W}\mathbf{x}^k - \gamma_k \mathbf{d}^k. \tag{15}$$

Our theoretical results rely on the following assumption.

Assumption 5: For any $\xi \in \Xi$, it holds that

1) $\|\nabla F(\mathbf{x}; \xi) - \nabla F(\mathbf{y}; \xi)\| \le L\|\mathbf{x} - \mathbf{y}\|$, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$;
2) $\sup_{\mathbf{x}\in\mathbb{R}^n, \xi\in\Xi}\{\|\nabla F(\mathbf{x}; \xi)\|\} < +\infty$;
3) $\mathbb{E}_\xi\left(\nabla F(\mathbf{x}; \xi)\right) = \nabla\left(\mathbb{E}_\xi F(\mathbf{x}; \xi)\right)$, $\forall \mathbf{x} \in \mathbb{R}^n$;
4) $\inf_{\mathbf{x}\in\mathbb{R}^n}\left(\mathbb{E}_{\xi_i} F(\mathbf{x}; \xi_i)\right) > -\infty$, $i = 1, 2, \ldots, m$;
5) The stationary distribution of the Markov chain in the i-th node is Π_i.

We have the following convergence result of DMGD for solving Problem (1) when all the underlying distributions are continuous.

Proposition 2: Let Assumption 5 hold, and $(\mathbf{x}^k)_{k\ge 0}$ denote the iterates generated by Equation (2) with Markov chain sampling, and condition Equation (14) hold. If the stepsizes are selected as Equation (9), we have the following convergence result

$$\lim_k \mathbb{E}\|\nabla \mathcal{F}(\overline{\mathbf{x}^k})\| = 0, \tag{16}$$

where $\mathcal{F}(\mathbf{x}) := \frac{1}{m}\sum_{i=1}^m \mathbb{E}_{\xi_i}\big(F(\mathbf{x}; \xi_i)\big)$. Moreover, the ergodic convergence rate is given by

$$\min_{1\le k\le K} \mathbb{E}\|\nabla \mathcal{F}(\overline{\mathbf{x}^k})\|^2 = \mathcal{O}\left(\frac{1 + \frac{1}{\ln^2(1/\lambda)}\cdot[1 + \frac{1}{\sqrt{m}(1-\lambda(\mathbf{W}))}] + \mathcal{F}(\overline{\mathbf{x}^0}) - \min \mathcal{F}}{K^{1-\theta}}\right).$$

Unlike Theorem 1, the Markov chain assumption cannot be weakened, i.e., the Markov chain must be time-homogeneous and reversible in Proposition 2. Another difference is that the stationary distributions $\Pi_1, \Pi_2, \ldots, \Pi_m$ are not necessarily to be uniform in Proposition 2.

V. Numerical Results

In this section, we compare the numerical performance of our proposed algorithm with DSGD on an autoregressive model, which closely resembles the first experiment in [53]. Assume that there are m autoregressive processes distributed on a graph of m nodes. We attempt to recover a consensus vector $\mathbf{u}$ from the multiple processes. On each node j, set matrix $\mathbf{A}^j$ as a subdiagonal matrix with random entries $\mathbf{A}_{l,l-1}(i) \overset{\text{i.i.d}}{\sim} \mathcal{U}[0.8, 0.99]$. Randomly sample a vector $\mathbf{u} \in \mathbb{R}^n$, with the unit 2-norm. In our experiments, we tested with $m = 50, n = 2000$. We want to numerically demonstrate the advantage of DMGD over DSGD-T in different decentralized topologies (connected graph). In particular, we select three classical graphs, including "cycle", "random" and "bipartite", all frequently used in decentralized experiments [31], [30], [59]. The "cycle" graph merely connects all nodes by a circle [60]; A "random" graph is more complicated, in which properties such as the number of graph vertices, graph edges, and connections between them are determined in some random way [61]; The "bipartite" graph is a graph in which every edge connects a vertex of one set to a vertex of the other set [62]. The data $(\xi_t^1(i), \xi_t^2(i))_{t=1}^\infty$ are generated by the following autoregressive process:

$$\xi_t^1(i) = \mathbf{A}(i)\xi_{t-1}^1(i) + \mathbf{e}_1 w^t,\ w^t \overset{\text{i.i.d}}{\sim} N(0,1),$$
$$\bar{\xi}_t^2(i) = \begin{cases} 1, & \text{if } \langle \mathbf{u}, \xi_t^1(i)\rangle > 0, \\ 0, & \text{otherwise;} \end{cases}$$
$$\xi_t^2(i) = \begin{cases} \bar{\xi}_t^2(i), & \text{with probability 0.8,} \\ 1 - \bar{\xi}_t^2(i), & \text{with probability 0.2.} \end{cases}$$

Clearly, for any $i \in \{1, 2, \ldots, M\}$, $(\xi_t^1(i), \xi_t^2(i))_{t=1}^\infty$ forms a Markov chain. Let Π_i denote the stationary distribution of the Markov chain on the i-th node. Three kinds of loss functions are used, which are given as follows

1) $\ell(\mathbf{x}; \xi^1(i), \xi^2(i)) = -\xi^2(i)\log(\sigma(\langle \mathbf{x}, \xi^1(i)\rangle)) - (1 - \xi^2(i))\log(1 - \sigma(\langle \mathbf{x}, \xi^1(i)\rangle))$,
2) $\ell(\mathbf{x}; \xi^1(i), \xi^2(i)) = -\xi^2(i)\log(\sigma(\langle \mathbf{x}, \xi^1(i)\rangle)) - (1 - \xi^2(i))\log(1 - \sigma(\langle \mathbf{x}, \xi^1(i)\rangle)) + 10^{-3}/2\|\mathbf{x}\|^2$,
3) $\ell(\mathbf{x}; \xi^1(i), \xi^2(i)) = 1/2\|\sigma(\langle \mathbf{x}, \xi^1(i)\rangle) - \xi^2(i)\|^2$,

where $\sigma(t) = \frac{1}{1+\exp(-t)}$. We reconstruct $\mathbf{x}$ as the solution to the following problem:

$$\min_{\mathbf{x}} \quad \sum_{i=1}^m \mathbb{E}_{(\xi^1(i), \xi^2(i))\sim\Pi_i} \ell(\mathbf{x}; \xi^1(i), \xi^2(i)). \tag{17}$$

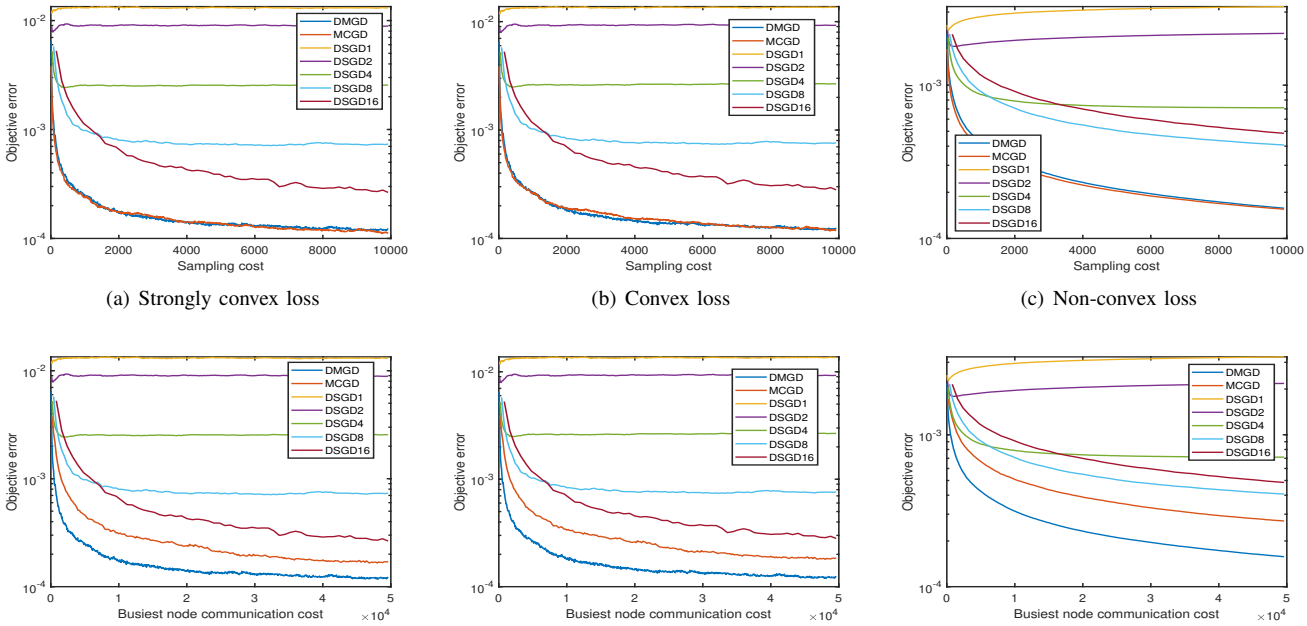
We choose $\gamma_k = \frac{1}{(k+1)^q}$ as our stepsize, where $q = 0.51$. This choice is consistent with our theory below. Specifically, we compare:

DMGD, where $(\xi^{1,k}(i), \xi^{2,k}(i))$ is from one trajectory of the Markov chain on the i-th node;

MCGD, (i.e., the centralized Markov chain gradient descent), where $(\xi^{1,k}(i), \xi^{2,k}(i))$ is from one trajectory of the Markov chain on the i-th worker;

DSGD-T, for $T = 1, 2, 4, 8, 16$, where each $(\xi^{1,k}(i), \xi^{2,k}(i))$ is the T-th sample of an independent trajectory on the i-th node. All trajectories are generated by starting from the same initial state.

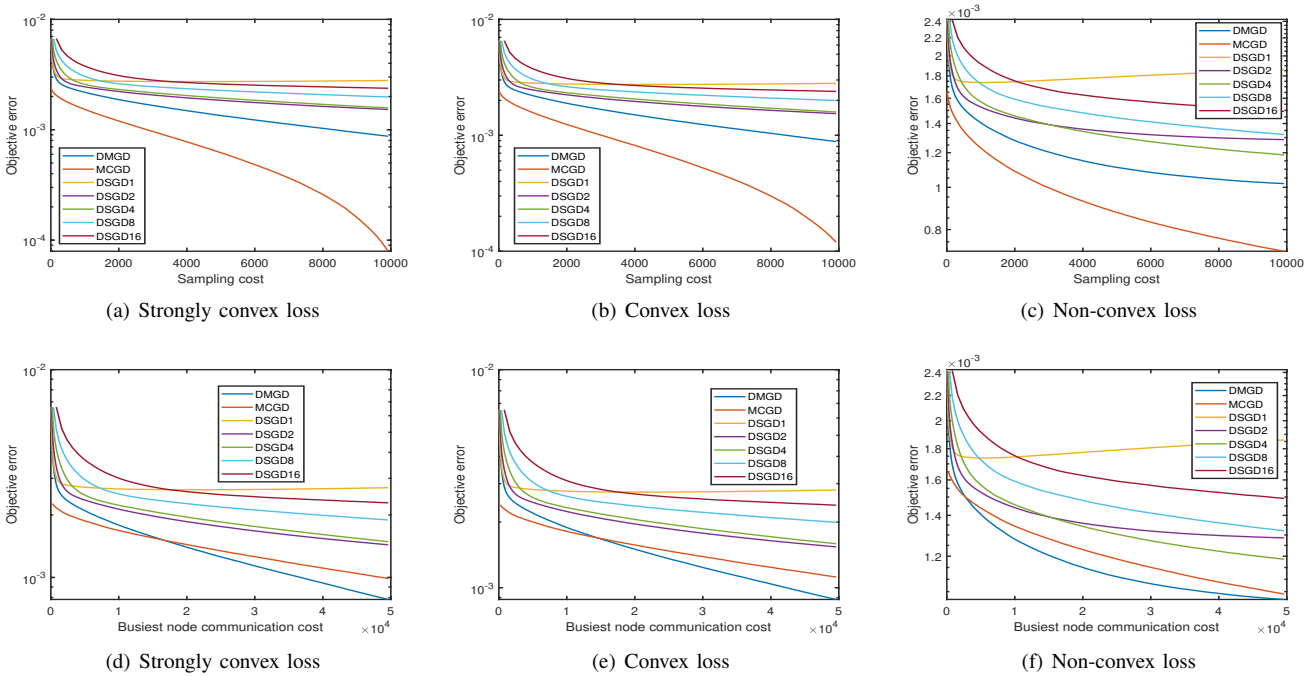
To compute T gradients, DSGD-T uses T times as many samples as DMGD. The sampling cost of DMGD-T collects all historical sampling associated with the gradients calculated. We did not try to adapt T as k increases because there is a lack of theoretical guidance. The communication cost is counted as $K \cdot \sum_{i=1}^m d_i$, where K is the iteration number and d_i is the degree of node i. All plots in the numerical tests are averaged over 5 rounds of simulations. The numerical comparisons are reported in Figures 1-3, which show that DMGD outperforms the DSGD-T with $T = 1, 2, 4, 8, 16$. The numerical results in Figures 1-3 are quite positive on DMGD. As expected, DMGD uses significantly fewer total samples than DSGD on each T. Surprisingly, DMGD did not cost even more gradient computations. It is important to note that DSGD-1 and DSGD-2, as well as DSGD-4, stagnate at noticeably lower accuracies due to their T values being too small. On the other hand, we observe that DMGD may not beat the centralized MCGD



(a) Strongly convex loss (b) Convex loss (c) Non-convex loss

(d) Strongly convex loss (e) Convex loss (f) Non-convex loss

Fig. 1. Comparisons of DMGD, centralized MCGD, and DSGD-T for $T = 1, 2, 4, 8, 16$ with "Random" graph.



(a) Strongly convex loss (b) Convex loss (c) Non-convex loss

(d) Strongly convex loss (e) Convex loss (f) Non-convex loss

Fig. 2. Comparisons of DMGD, centralized MCGD, and DSGD-T for $T = 1, 2, 4, 8, 16$ with "Cycle" graph.

in sampling costs, but it can significantly reduce the busiest node's communications.

VI. Proofs

A. Technical lemmas

Lemma 1 (Lemma 1, [18]): Let Assumption 4 hold, and $\lambda_i(\mathbf{H}) \in \mathbb{C}$ be the i-th largest eigenvalue of $\mathbf{H} \in \mathbb{R}^{d\times d}$, and $\lambda(\mathbf{H}) := \frac{\max\{|\lambda_2(\mathbf{H})|, |\lambda_{\min}(\mathbf{H})|\}+1}{2} \in [0, 1)$. Then, we can bound the largest entry-wise absolute value of the deviation matrix $\delta^k := \Pi^* - \mathbf{H}^k \in \mathbb{R}^{d\times d}$ as

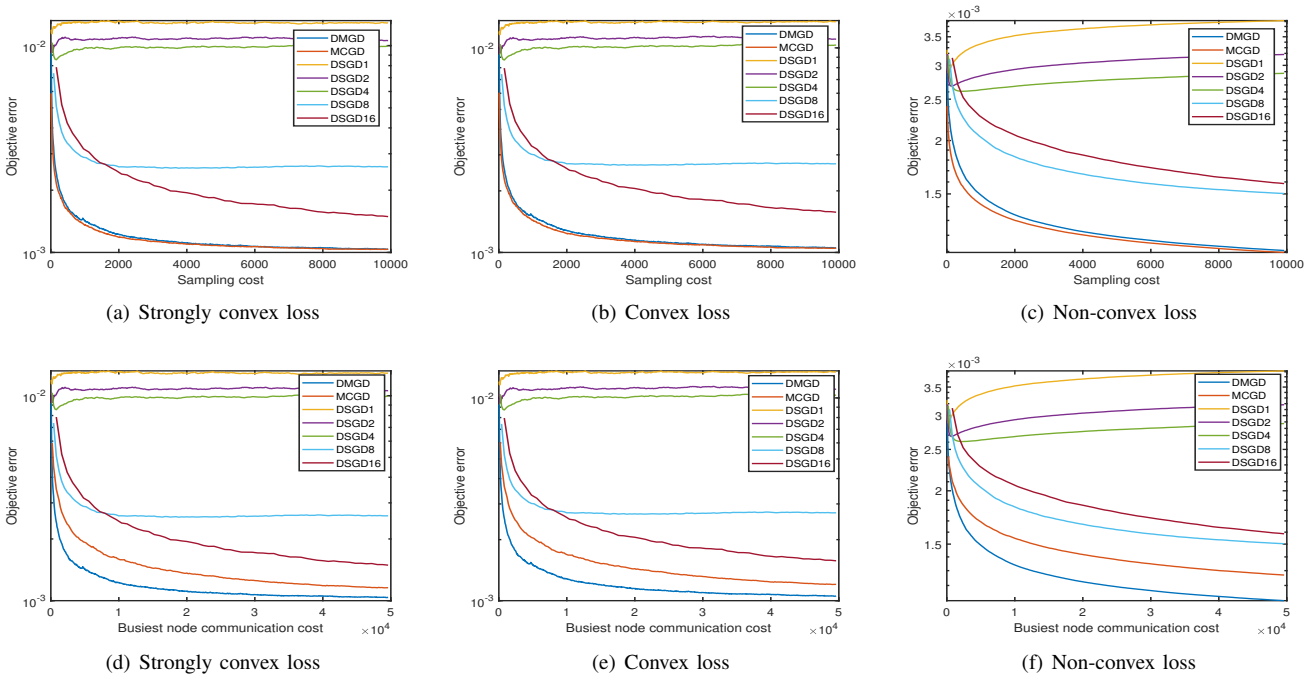
$$\|\delta^k\|_\infty \leq C_{\mathbf{H}} \cdot [\lambda(\mathbf{H})]^k \tag{18}$$

as $k \geq K_{\mathbf{H}}$, where $C_{\mathbf{H}}$ is a constant that depends on the Jordan canonical form of $\mathbf{H}$ and $K_{\mathbf{H}}$ is a constant that depends on $\lambda_2(\mathbf{H})$ and $\lambda_{\min}(\mathbf{H})$.

Lemma 2 (Corollary 1.14., [17]): Let $\mathbf{P} \in \mathbb{R}^{m\times m}$ be the matrix whose elements are all $1/m$. Given any $k \in \mathbb{Z}^+$, the mixing matrix $\mathbf{W} \in \mathbb{R}^{m\times m}$ satisfies

$$\|\mathbf{W}^k - \mathbf{P}\| \leq [\lambda(\mathbf{W})]^k.$$

Lemma 3: Let $(\mathbf{x}^k)_{k\geq 0}$ be generated by DMGD and As-



(a) Strongly convex loss (b) Convex loss (c) Non-convex loss

(d) Strongly convex loss (e) Convex loss (f) Non-convex loss

Fig. 3. Comparisons of DMGD, centralized MCGD, and DSGD-T for $T = 1, 2, 4, 8, 16$ with "Bipartite" graph.

sumption 2 hold, then we have

$$\frac{1}{m}\sum_{i=1}^{m}\|\mathbf{x}^k(i)-\overline{\mathbf{x}^k}\| \le B\sum_{j=0}^{k}\gamma_j[\lambda(\mathbf{W})]^{k-j}. \tag{19}$$

for any $k \ge 0$. Further if $\gamma_j = \frac{1}{(j+1)^\theta}$ with $\frac{1}{2} \le \theta < 1$, it follows

$$\frac{1}{m}\sum_{i=1}^{m}\|\mathbf{x}^k(i)-\overline{\mathbf{x}^k}\| \le \frac{BC_{\mathbf{W}}}{(k+1)^\theta} = BC_{\mathbf{W}}\gamma_k, \tag{20}$$

and $C_{\mathbf{W}}$ is a positive constant dependent on $\mathbf{W}$, and $C_{\mathbf{W}} = \mathcal{O}(\frac{1}{1-\lambda(\mathbf{W})})$.

Lemma 4 (Lemma 2, [18]): Consider two nonnegative sequences $(\beta_k)_{k\ge0}$ and $(h_k)_{k\ge0}$ that satisfy 1) $\lim_k h_k = 0$ and $\sum_k h_k = +\infty$, 2) $\sum_k \beta_k h_k < +\infty$, 3) $|\beta_{k+1}-\beta_k| \le ch_k$ for some $c > 0$ and $k = 0, 1, \ldots$. Then, we have $\lim \beta_k = 0$.

Lemma 5: Let $a > 0$, c be a real number, and $x \ge \max\{|a(\ln 2a - 1) - c|, e^{\frac{|c|}{a}}/4, 16\}$. If $y - a\ln y + c = x$, it then holds $y - x \le 3a\ln x$.

Lemma 6: Given any fixed positive number $a > 0$, it holds $t \le \frac{a}{e}e^{\frac{t}{a}}$, where $t > 0$.

B. Proof of Theorem 1

The proof consists of four major steps: 1. Introduce the delay $\mathcal{T}_k$ and its property. 2. Bound $\sum_k \gamma_k\mathbb{E}\|\nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}})\|^2$ by four terms. 3. Prove the upper bound of the sum of the four terms. 4. Establish the upper bound for $\sum_k \big(\gamma_k\mathbb{E}\|\nabla f(\overline{\mathbf{x}^k})\|^2 - \gamma_k\mathbb{E}\|\nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}})\|^2\big)$.

Step 1: For integer $k \ge 1$, denote the integer $\mathcal{T}_k$ as

$$\mathcal{T}_k := \min\{\max\left\{\left\lceil \ln\left(2MC_{\mathbf{H}}B^2k\right)\Big/\ln\frac{1}{\lambda(\mathbf{H})}\right\rceil, K_{\mathbf{H}}\right\}, k\},$$

where $K_{\mathbf{H}}$ and $\lambda(\mathbf{H})$ are defined in Lemma 1. We can see that $\mathcal{T}_k \le k$ for any $k \in \mathbb{Z}^+$.

Because we need to prove the upper bound, which is still finite when some fixed terms are removed. We can choose k that is sufficiently large. Thus, in the following part, there are some arguments of the form of "as $k \ge c$". Indeed, we want to determine the constant c, which can guarantee the upper bounds to be finite in the four steps when $k \ge c$ (see Equation (37)).

When $k \ge \frac{[\frac{1}{\lambda(\mathbf{H})}]^{K_{\mathbf{H}}}}{2MC_{\mathbf{H}}B^2}$, we have

$$\left\lceil\frac{\ln\left(2MC_{\mathbf{H}}B^2k\right)}{\ln\frac{1}{\lambda(\mathbf{H})}}\right\rceil \ge \ln\left(2MC_{\mathbf{H}}B^2k\right)\Big/\ln\frac{1}{\lambda(\mathbf{H})} \ge K_{\mathbf{H}},$$

and $\mathcal{T}_k = \min\{\left\lceil \ln(2MC_{\mathbf{H}}B^2k)\Big/\ln\frac{1}{\lambda(\mathbf{H})}\right\rceil, k\}$ in this case. In Equation (53), letting $t \leftarrow k$ and $a \leftarrow \frac{2}{\ln\frac{1}{\lambda(\mathbf{H})}}$, we can get

$$\ln k \le \frac{k\ln\frac{1}{\lambda(\mathbf{H})}}{2} - \ln\ln\frac{1}{\lambda(\mathbf{H})} + \ln\frac{2}{e} \le \frac{k\ln\frac{1}{\lambda(\mathbf{H})}}{2} - \ln\ln\frac{1}{\lambda(\mathbf{H})},$$

where we used the fact that $\ln\frac{2}{e} < 0$. Due to $0 < \lambda(\mathbf{H}) < 1$ and $\ln\frac{1}{\lambda(\mathbf{H})} > 0$, we then get $\frac{\ln k}{\ln\frac{1}{\lambda(\mathbf{H})}} \le \frac{k}{2} - \frac{\ln\ln\frac{1}{\lambda(\mathbf{H})}}{\ln\frac{1}{\lambda(\mathbf{H})}}$, based on which we get $\frac{\ln(2MC_{\mathbf{H}}B^2k)}{\ln\frac{1}{\lambda(\mathbf{H})}} \le \frac{k}{\ln\frac{1}{\lambda(\mathbf{H})}} + \frac{\ln 2MC_{\mathbf{H}}B^2}{\ln\frac{1}{\lambda(\mathbf{H})}} \le \frac{k}{2} - \frac{\ln\ln\frac{1}{\lambda(\mathbf{H})}}{\ln\frac{1}{\lambda(\mathbf{H})}} + \frac{\ln 2MC_{\mathbf{H}}B^2}{\ln\frac{1}{\lambda(\mathbf{H})}} \le k$, provided $k \ge \frac{2\ln\frac{2MC_{\mathbf{H}}B^2}{\ln\frac{1}{\lambda(\mathbf{H})}}}{\ln\frac{1}{\lambda(\mathbf{H})}}$. In summary, when

$$k \ge \max\left\{\frac{[\frac{1}{\lambda(\mathbf{H})}]^{K_{\mathbf{H}}}}{2MC_{\mathbf{H}}B^2}, \frac{2\ln\frac{2MC_{\mathbf{H}}B^2}{\ln\frac{1}{\lambda(\mathbf{H})}}}{\ln\frac{1}{\lambda(\mathbf{H})}}, K_{\mathbf{H}}\right\}, \tag{21}$$

we have $\mathcal{T}_k = \left\lceil \ln(2MC_{\mathbf{H}}B^2k)\Big/\ln\frac{1}{\lambda(\mathbf{H})}\right\rceil$. Assume (21) holds, in this case notice that $\mathcal{T}_k \ge K_{\mathbf{H}}$ as $k \ge K_{\mathbf{H}}$, by

using Lemma 1, we then get

$$\begin{aligned}&\left|[\mathbf{H}^{\mathcal{T}_k}]_{i,j}-\frac{1}{M}\right|\le\|\delta^{\mathcal{T}_k}\|_\infty\le C_{\mathbf{H}}\cdot[\lambda(\mathbf{H})]^{\mathcal{T}_k}\\&\overset{a)}{\le}\max\{\min\{\frac{1/k}{2MB^2},C_{\mathbf{H}}\cdot[\lambda(\mathbf{H})]^{K_{\mathbf{H}}}\},C_{\mathbf{H}}\cdot[\lambda(\mathbf{H})]^k\},\end{aligned}\tag{22}$$

for any $i,j\in\{1,2,\ldots,M\}$, where δ^k, $C_{\mathbf{H}}$ and $K_{\mathbf{H}}$ are given in Lemma 1, and $a)$ uses the following result

$$\begin{aligned}&C_{\mathbf{H}}\cdot[\lambda(\mathbf{H})]^{\left\lceil\ln(2MC_{\mathbf{H}}B^2k)\Big/\ln\frac{1}{\lambda(\mathbf{H})}\right\rceil}\\&\le C_{\mathbf{H}}\cdot[\lambda(\mathbf{H})]^{\ln(2MC_{\mathbf{H}}B^2k)\Big/\ln\frac{1}{\lambda(\mathbf{H})}}\\&=C_{\mathbf{H}}\cdot[\exp(-\ln\frac{1}{\lambda(\mathbf{H})})]^{\ln(2MC_{\mathbf{H}}B^2k)\Big/\ln\frac{1}{\lambda(\mathbf{H})}}\\&=C_{\mathbf{H}}\cdot[\exp(-\ln\frac{1}{\lambda(\mathbf{H})}\cdot\ln(2MC_{\mathbf{H}}B^2k)\Big/\ln\frac{1}{\lambda(\mathbf{H})})]\\&=C_{\mathbf{H}}\cdot\exp(\ln(\frac{1}{2MC_{\mathbf{H}}B^2k}))=\frac{1}{2MB^2k}.\end{aligned}$$

Furthermore, if $k\ge K_{\mathbf{H}}$, the rightmost part of Equation (22) can be bounded by $\max\{\frac{1/k}{2MB^2},C_{\mathbf{H}}\cdot[\lambda(\mathbf{H})]^k\}$. In Equation (53), by letting $t\leftarrow k$ and $a\leftarrow\frac{1}{\ln\frac{1}{\lambda(\mathbf{H})}}$, we obtain $k\ln\lambda(\mathbf{H})+\ln\ln\frac{1}{\lambda(\mathbf{H})}+1\le-\ln k$. Adding $\ln C_{\mathbf{H}}$ to both sides and taking the exponential, we have $C_{\mathbf{H}}\cdot[\lambda(\mathbf{H})]^k\le\frac{C_{\mathbf{H}}}{e\ln\frac{1}{\lambda(\mathbf{H})}}\frac{1}{k}$. Note that k is unnecessary to be an integer here. By replacing $k\leftarrow k/2$, we get

$$C_{\mathbf{H}}\cdot[\lambda(\mathbf{H})]^{k/2}\le\frac{2C_{\mathbf{H}}}{e\ln\frac{1}{\lambda(\mathbf{H})}}\frac{1}{k}.\tag{23}$$

With Equation (23), we can further get

$$\begin{aligned}C_{\mathbf{H}}\cdot[\lambda(\mathbf{H})]^k&\le C_{\mathbf{H}}\cdot[\lambda(\mathbf{H})]^{k/2}\cdot[\lambda(\mathbf{H})]^{k/2}\\&\le\Big([\lambda(\mathbf{H})]^{k/2}\frac{2C_{\mathbf{H}}}{e\ln\frac{1}{\lambda(\mathbf{H})}}\Big)\frac{1}{k}.\end{aligned}\tag{24}$$

When $k\ge\frac{2\ln\frac{4MC_{\mathbf{H}}B^2}{e\ln\frac{1}{\lambda(\mathbf{H})}}}{\ln\frac{1}{\lambda(\mathbf{H})}}$, we have

$$[\lambda(\mathbf{H})]^{k/2}\frac{2C_{\mathbf{H}}}{e\ln\frac{1}{\lambda(\mathbf{H})}}\le\frac{1}{2MB^2}\overset{(24)}{\Longrightarrow}C_{\mathbf{H}}\cdot[\lambda(\mathbf{H})]^k\le\frac{1/k}{2MB^2}.$$

Note that Equation (24) is established based on the fact that $k\ge K_{\mathbf{H}}$ and $k\ge1$. Thus, when

$$k\ge\max\left\{\frac{2\ln\frac{4MC_{\mathbf{H}}B^2}{e\ln\frac{1}{\lambda(\mathbf{H})}}}{\ln\frac{1}{\lambda(\mathbf{H})}},K_{\mathbf{H}},1\right\},\tag{25}$$

we have $C_{\mathbf{H}}\cdot[\lambda(\mathbf{H})]^k\le\frac{1/k}{2MB^2}$, and

$$\left|[\mathbf{H}^{\mathcal{T}_k}]_{i,j}-\frac{1}{M}\right|\le\frac{1/k}{2MB^2}.\tag{26}$$

Step 2. Denote the shorthand notations $\tilde{\mathbf{u}}^k:=\frac{\sum_{h=1}^m\nabla f_h^{j_{h,k}}(\mathbf{x}^k(h))}{m}$, $\tilde{\mathbf{u}}^{k-\mathcal{T}_k}:=\frac{\sum_{h=1}^m\nabla f_h^{j_{h,k}}(\mathbf{x}^{k-\mathcal{T}_k}(h))}{m}$, when Equation (25) holds, we have

$$\begin{aligned}&\mathbb{E}_{\mathbf{j}_k}(\tilde{\mathbf{u}}^{k-\mathcal{T}_k}\mid\chi^{k-\mathcal{T}_k})\\&\overset{a)}{=}\frac{1}{m}\sum_{h=1}^m\sum_{i=1}^M\nabla f_h^{j_{h,k}}(\mathbf{x}^{k-\mathcal{T}_k}(h))\cdot\mathbb{P}(j_{h,k}=i\mid\chi^{k-\mathcal{T}_k})\Big\rangle\\&\overset{b)}{=}\frac{1}{m}\sum_{h=1}^m\sum_{i=1}^M\nabla f_h^{j_{h,k}}(\mathbf{x}^{k-\mathcal{T}_k}(h))\cdot\mathbb{P}(j_{h,k}=i\mid j_{h,k-\mathcal{T}_k})\\&\overset{c)}{=}\frac{1}{m}\sum_{h=1}^m\sum_{i=1}^M\nabla f_h^i(\mathbf{x}^{k-\mathcal{T}_k}(h))\cdot[\mathbf{H}^{\mathcal{T}_k}]_{j_{h,k-\mathcal{T}_k},i},\end{aligned}\tag{27}$$

where $a)$ is due to the conditional expectation, and $b)$ uses the property of Markov chain, and $c)$ is the matrix form of the probability. On the other hand, we are led to the following estimate

$$\begin{aligned}&\mathbb{E}_{\mathbf{j}_k}(\langle\nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}}),\tilde{\mathbf{u}}^{k-\mathcal{T}_k}\rangle\mid\chi^{k-\mathcal{T}_k})\\&=\left\langle\nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}}),\frac{1}{m}\sum_{h=1}^m\frac{1}{M}\sum_{i=1}^M\nabla f_h^i(\mathbf{x}^{k-\mathcal{T}_k}(h))\right\rangle\\&+\left\langle\nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}}),\frac{1}{m}\sum_{h=1}^m\sum_{i=1}^M\nabla f_h^i(\mathbf{x}^{k-\mathcal{T}_k}(h))\right.\\&\quad\left.\cdot\Big[[\mathbf{H}^{\mathcal{T}_k}]_{j_{h,k-\mathcal{T}_k},i}-\frac{1}{M}\Big]\right\rangle\\&\overset{\text{Equation (26)}}{\ge}\left\langle\nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}}),\frac{1}{m}\sum_{i=1}^m\nabla f_i(\mathbf{x}^{k-\mathcal{T}_k}(i))\right\rangle\\&-\|\nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}})\|\cdot\frac{1}{m}\sum_{h=1}^m\sum_{i=1}^M\|\nabla f_h^i(\mathbf{x}^{k-\mathcal{T}_k}(h))\|\cdot\frac{1/k}{2MB^2}\\&\ge\left\langle\nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}}),\frac{1}{m}\sum_{i=1}^m\nabla f_i(\mathbf{x}^{k-\mathcal{T}_k}(i))\right\rangle-\frac{1}{2k}.\end{aligned}$$

Direct calculations yield

$$\begin{aligned}&\left\langle\nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}}),\frac{1}{m}\sum_{i=1}^m\nabla f_i(\mathbf{x}^{k-\mathcal{T}_k}(i))\right\rangle=\|\nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}})\|^2\\&+\langle\nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}}),\frac{1}{m}\sum_{i=1}^m\nabla f_i(\mathbf{x}^{k-\mathcal{T}_k}(i))-\nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}})\rangle\\&\ge\|\nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}})\|^2-B\frac{L}{m}\sum_{i=1}^m\|\mathbf{x}^{k-\mathcal{T}_k}(i)-\overline{\mathbf{x}^{k-\mathcal{T}_k}}\|,\end{aligned}\tag{28}$$

where the last inequality uses Assumptions 2 and 3. Rearrangement of Equation (27) together with Equation (28) gives us

$$\begin{aligned}\gamma_k\mathbb{E}\|\nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}})\|^2&\le\gamma_k\mathbb{E}(\langle\nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}}),\tilde{\mathbf{u}}^{k-\mathcal{T}_k}\rangle)\\&+B\frac{L}{m}\sum_{i=1}^m\gamma_k\mathbb{E}\|\mathbf{x}^{k-\mathcal{T}_k}(i)-\overline{\mathbf{x}^{k-\mathcal{T}_k}}\|+\frac{\gamma_k}{2k}.\end{aligned}\tag{29}$$

We can bound $f(\overline{\mathbf{x}^{k+1}}) - f(\overline{\mathbf{x}^k})$ as follows

$$
\begin{aligned}
&f(\overline{\mathbf{x}^{k+1}}) - f(\overline{\mathbf{x}^k}) \overset{a)}{\leq} \langle \nabla f(\overline{\mathbf{x}^k}), \overline{\mathbf{x}^{k+1}} - \overline{\mathbf{x}^k} \rangle \\
&+ \frac{L}{2}\|\overline{\mathbf{x}^{k+1}} - \overline{\mathbf{x}^k}\|^2 \overset{b)}{=} \langle \nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}}), \overline{\mathbf{x}^{k+1}} - \overline{\mathbf{x}^k} \rangle \\
&+ \langle \nabla f(\overline{\mathbf{x}^k}) - \nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}}), \overline{\mathbf{x}^{k+1}} - \overline{\mathbf{x}^k} \rangle + \frac{L}{2}\|\overline{\mathbf{x}^{k+1}} - \overline{\mathbf{x}^k}\|^2 \\
&\overset{c)}{\leq} \langle \nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}}), \overline{\mathbf{x}^{k+1}} - \overline{\mathbf{x}^k} \rangle \\
&+ \frac{(L+1)\|\overline{\mathbf{x}^{k+1}} - \overline{\mathbf{x}^k}\|^2}{2} + \frac{L^2\|\overline{\mathbf{x}^k} - \overline{\mathbf{x}^{k-\mathcal{T}_k}}\|^2}{2},
\end{aligned} \tag{30}
$$

where $a)$ depends on the continuity of ∇f, and $b)$ is the basic algebraic computation, $c)$ uses Schwarz inequality $\langle \nabla f(\overline{\mathbf{x}^k}) - \nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}}), \overline{\mathbf{x}^{k+1}} - \overline{\mathbf{x}^k} \rangle \leq L\|\mathbf{x}^k - \mathbf{x}^{k-\mathcal{T}_k}\| \cdot \|\overline{\mathbf{x}^{k+1}} - \overline{\mathbf{x}^k}\| \leq \frac{L^2\|\mathbf{x}^k - \mathbf{x}^{k-\mathcal{T}_k}\|^2}{2} + \frac{\|\overline{\mathbf{x}^{k+1}} - \overline{\mathbf{x}^k}\|^2}{2}$. Moving $\langle \nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}}), \overline{\mathbf{x}^{k+1}} - \overline{\mathbf{x}^k} \rangle$ to left-hand side, we have

$$
\begin{aligned}
&\langle \nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}}), \overline{\mathbf{x}^k} - \overline{\mathbf{x}^{k+1}} \rangle \leq f(\overline{\mathbf{x}^k}) - f(\overline{\mathbf{x}^{k+1}}) \\
&+ \frac{L^2\|\overline{\mathbf{x}^{k+1}} - \overline{\mathbf{x}^{k-\mathcal{T}_k}}\|^2}{2} + \frac{(L+1)\|\overline{\mathbf{x}^{k+1}} - \overline{\mathbf{x}^k}\|^2}{2}.
\end{aligned} \tag{31}
$$

We then consider the following bound:

$$
\begin{aligned}
&\mathbb{E}(\langle \nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}}), \overline{\mathbf{x}^k} - \overline{\mathbf{x}^{k+1}} \rangle \mid \chi^{k-\mathcal{T}_k}) \\
&= \gamma_k \mathbb{E}(\langle \nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}}), \tilde{\mathbf{u}}^{k-\mathcal{T}_k} \rangle \mid \chi^{k-\mathcal{T}_k}) \\
&\quad + \gamma_k \mathbb{E}(\langle \nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}}), \tilde{\mathbf{u}}^k - \tilde{\mathbf{u}}^{k-\mathcal{T}_k} \rangle \mid \chi^{k-\mathcal{T}_k}) \\
&\geq \gamma_k \mathbb{E}(\langle \nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}}), \tilde{\mathbf{u}}^{k-\mathcal{T}_k} \rangle \mid \chi^{k-\mathcal{T}_k}) \\
&\quad - BL \cdot \mathbb{E}(\gamma_k \|\overline{\mathbf{x}^k} - \overline{\mathbf{x}^{k-\mathcal{T}_k}}\| \mid \chi^{k-\mathcal{T}_k}) \\
&\quad - \gamma_k BL \mathbb{E}\Big(\frac{\sum_{h=1}^m \|\mathbf{x}^k(h) - \overline{\mathbf{x}^k}\|}{m} \\
&\quad + \frac{\sum_{h=1}^m \|\mathbf{x}^{k-\mathcal{T}_k}(h) - \overline{\mathbf{x}^{k-\mathcal{T}_k}}\|}{m} \mid \chi^{k-\mathcal{T}_k} \Big)
\end{aligned} \tag{32}
$$

where the last inequality uses the following estimate

$$
\begin{aligned}
&\gamma_k \mathbb{E}(\langle \nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}}), \tilde{\mathbf{u}}^k - \tilde{\mathbf{u}}^{k-\mathcal{T}_k} \rangle \mid \chi^{k-\mathcal{T}_k}) \\
&\geq -\gamma_k \mathbb{E}(\|\nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}})\| \cdot \|\tilde{\mathbf{u}}^k - \tilde{\mathbf{u}}^{k-\mathcal{T}_k}\| \mid \chi^{k-\mathcal{T}_k}) \\
&\geq -\gamma_k BL \mathbb{E}\Big(\frac{\sum_{h=1}^m \|\mathbf{x}^k(h) - \mathbf{x}^{k-\mathcal{T}_k}(h)\|}{m} \mid \chi^{k-\mathcal{T}_k} \Big) \\
&\geq -\gamma_k BL \mathbb{E}\Big(\frac{\sum_{h=1}^m \|\mathbf{x}^k(h) - \overline{\mathbf{x}^k}\|}{m} \\
&\quad + \|\overline{\mathbf{x}^k} - \overline{\mathbf{x}^{k-\mathcal{T}_k}}\| + \frac{\sum_{h=1}^m \|\overline{\mathbf{x}^{k-\mathcal{T}_k}} - \mathbf{x}^{k-\mathcal{T}_k}(h)\|}{m} \mid \chi^{k-\mathcal{T}_k} \Big)
\end{aligned}
$$

with the Lipschitz continuity of $\nabla f_h^{j_{h,k}}$ and boundedness of ∇f. Taking conditional expectations on both sides of Equation (31) on $\chi^{k-\mathcal{T}_k}$ and rearrangement of Equation (32), then we have

$$
\begin{aligned}
&\gamma_k \mathbb{E}_{\mathbf{j}_k}(\langle \nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}}), \tilde{\mathbf{u}}^{k-\mathcal{T}_k} \rangle \mid \chi^{k-\mathcal{T}_k}) \\
&\leq \mathbb{E}\big(f(\overline{\mathbf{x}^k}) - f(\overline{\mathbf{x}^{k+1}}) \mid \chi^{k-\mathcal{T}_k}\big) \\
&\quad + \frac{(L+1) \cdot \mathbb{E}(\|\overline{\mathbf{x}^{k+1}} - \overline{\mathbf{x}^k}\|^2 \mid \chi^{k-\mathcal{T}_k})}{2} \\
&\quad + BL \cdot \mathbb{E}(\gamma_k \|\overline{\mathbf{x}^k} - \overline{\mathbf{x}^{k-\mathcal{T}_k}}\| \mid \chi^{k-\mathcal{T}_k}) \\
&\quad + \frac{L^2 \cdot \mathbb{E}(\|\overline{\mathbf{x}^{k+1}} - \overline{\mathbf{x}^{k-\mathcal{T}_k}}\|^2 \mid \chi^{k-\mathcal{T}_k})}{2} \\
&\quad + \gamma_k BL \mathbb{E}\Big(\frac{\sum_{h=1}^m \|\overline{\mathbf{x}^k} - \mathbf{x}^k(h)\|}{m} \mid \chi^{k-\mathcal{T}_k} \Big) \\
&\quad + \gamma_k BL \mathbb{E}\Big(\frac{\sum_{h=1}^m \|\overline{\mathbf{x}^{k-\mathcal{T}_k}} - \mathbf{x}^{k-\mathcal{T}_k}(h)\|}{m} \mid \chi^{k-\mathcal{T}_k} \Big).
\end{aligned} \tag{33}
$$

Taking the expectation on both sides of Equation (33) on χ^k and with (29), using the total expectation rule $\mathbb{E}\Big[\mathbb{E}_{\mathbf{j}_k}(\langle \nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}}), \tilde{\mathbf{u}}^{k-\mathcal{T}_k} \rangle \mid \chi^{k-\mathcal{T}_k}) \mid \chi^k\Big] = \mathbb{E}(\langle \nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}}), \tilde{\mathbf{u}}^{k-\mathcal{T}_k} \rangle \mid \chi^k)$, we are then led to

$$
\gamma_k \mathbb{E}\|\nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}})\|^2 \leq \text{(I)} + \text{(II)} + \text{(III)} + \text{(IV)} + \text{(V)} + \frac{\gamma_k}{2k}, \tag{34}
$$

where (I) $:= \mathbb{E}\big(f(\overline{\mathbf{x}^k}) - f(\overline{\mathbf{x}^{k+1}})\big)$, (II) $:= \frac{(L+1)\cdot\mathbb{E}\|\overline{\mathbf{x}^{k+1}} - \overline{\mathbf{x}^k}\|^2}{2}$, (III) $:= BL\gamma_k \cdot \mathbb{E}\|\overline{\mathbf{x}^k} - \overline{\mathbf{x}^{k-\mathcal{T}_k}}\|$, (IV) $:= \frac{L^2 \cdot \mathbb{E}\|\overline{\mathbf{x}^{k+1}} - \overline{\mathbf{x}^{k-\mathcal{T}_k}}\|^2}{2}$, and (V) $:= 2B\frac{L}{m}\sum_{i=1}^m \gamma_k \mathbb{E}\|\mathbf{x}^{k-\mathcal{T}_k}(i) - \overline{\mathbf{x}^{k-\mathcal{T}_k}}\| + B\frac{L}{m}\sum_{i=1}^m \gamma_k \mathbb{E}\|\mathbf{x}^k(i) - \overline{\mathbf{x}^k}\|$.

Step 3: We now prove that (II), (III), (IV) and (V) are all summable. The summability (I) is obvious. For (II), (III) and (IV), with the fact that $\overline{\mathbf{x}^{k+1}} - \overline{\mathbf{x}^k} = -\gamma_k \frac{\sum_{i=1}^m \nabla f_i^{j_{i,k}}(\mathbf{x}^k(i))}{m}$, we can derive (we omit the constant hyper-parameters in following)

$$
\begin{aligned}
\text{(II)} : \mathbb{E}\|\overline{\mathbf{x}^{k+1}} - \overline{\mathbf{x}^k}\|^2 &= \gamma_k^2 \mathbb{E}\Big\| \sum_{i=1}^m \nabla f_i^{j_{i,k}}(\mathbf{x}^k(i))/m \Big\|^2 \\
&= \gamma_k^2 \sum_{i=1}^m \mathbb{E}\Big\| \nabla f_i^{j_{i,k}}(\mathbf{x}^k(i)) \Big\|^2 / m^2 \leq \gamma_k^2 B^2/m
\end{aligned}
$$

due to the dependence of $\{j_{i,k}\}_{1\leq i\leq m}$, and

$$
\begin{aligned}
\text{(III)} :&\mathbb{E}(\gamma_k \|\overline{\mathbf{x}^k} - \overline{\mathbf{x}^{k-\mathcal{T}_k}}\|) \leq \gamma_k \sum_{d=k-\mathcal{T}_k}^{k-1} \mathbb{E}\|\overline{\mathbf{x}^{d+1}} - \overline{\mathbf{x}^d}\| \\
&\leq \frac{B}{\sqrt{m}} \sum_{d=k-\mathcal{T}_k}^{k-1} \gamma_d \gamma_k \leq \frac{B}{2\sqrt{m}} \sum_{d=k-\mathcal{T}_k}^{k-1} (\gamma_d^2 + \gamma_k^2) \\
&= \frac{\mathcal{T}_k B}{2\sqrt{m}} \gamma_k^2 + \frac{B}{2\sqrt{m}} \sum_{d=k-\mathcal{T}_k}^{k-1} \gamma_d^2,
\end{aligned}
$$

and

$$
\begin{aligned}
&\text{(IV)} : \mathbb{E}(\|\overline{\mathbf{x}^{k+1}} - \overline{\mathbf{x}^{k-\mathcal{T}_k}}\|^2) \leq (\mathcal{T}_k + 1) \sum_{d=k-\mathcal{T}_k}^{k} \mathbb{E}\|\overline{\mathbf{x}^{d+1}} - \overline{\mathbf{x}^d}\|^2 \\
&\leq B^2(\mathcal{T}_k + 1)/m \sum_{d=k-\mathcal{T}_k}^{k} \gamma_d^2,
\end{aligned}
$$

and

$$\begin{aligned}
&\text{(V)} : 2BL\gamma_k(\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\|\mathbf{x}^{k-\mathcal{T}_k}(i)-\overline{\mathbf{x}^{k-\mathcal{T}_k}}\|)\\
&\quad + BL\gamma_k(\frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\|\mathbf{x}^{k}(i)-\overline{\mathbf{x}^{k}}\|)\\
&\overset{(20)}{\leq} 2B^2LC_{\mathbf{W}}\gamma_k\gamma_{k-\mathcal{T}_k}+B^2LC_{\mathbf{W}}\gamma_k^2 \leq B^2LC_{\mathbf{W}}(\gamma_{k-\mathcal{T}_k}^2+\gamma_k^2)\\
&+B^2LC_{\mathbf{W}}\gamma_k^2 \leq 2B^2LC_{\mathbf{W}}\sum_{d=k-\mathcal{T}_k}^{k}\gamma_d^2.
\end{aligned}$$

It is easy to see that if $(\mathcal{T}_k\sum_{d=k-\mathcal{T}_k}^{k}\gamma_d^2)_{k\geq 1}$ is summable, (II), (III) and (IV) are all summable. We consider the case

$$\begin{aligned}
k \geq K_0 := \max\Bigg\{&\frac{[\frac{1}{\lambda(\mathbf{H})}]^{K_{\mathbf{H}}}}{2MC_{\mathbf{H}}B^2}, \frac{2\ln\frac{2MC_{\mathbf{H}}B^2}{\ln\frac{1}{\lambda(\mathbf{H})}}}{\ln\frac{1}{\lambda(\mathbf{H})}}, K_{\mathbf{H}},\\
&\frac{2e\Big(\Big|\frac{\ln\ln\frac{1}{\lambda(\mathbf{H})}}{\ln\frac{1}{\lambda(\mathbf{H})}}\Big|+\Big|\frac{\ln 2MC_{\mathbf{H}}B^2}{\frac{1}{\lambda(\mathbf{H})}}\Big|\Big)}{e-2}\Bigg\},
\end{aligned} \tag{35}$$

which can let Lemma 1 be active, and $\mathcal{T}_k = \Big\lceil \ln(2MC_{\mathbf{H}}B^2k)\Big/\ln\frac{1}{\lambda(\mathbf{H})}\Big\rceil$. Note that the summability of sequence is free of finite items; we then consider $(\mathcal{T}_k\sum_{d=k-\mathcal{T}_k}^{k-1}\gamma_d^2)_{k\geq K_0}$. For any fixed integer $t \geq K_0$, γ_t^2 only appears at index $k \geq K_0$ satisfying $S_t := \{k\in\mathbb{Z}^+ \mid k-\mathcal{T}_k \leq t \leq k-1,\ k\geq K_0\}$ in the inner summation. Let $k = k(t)$ be the solution of $k - \ln(2MC_{\mathbf{H}}B^2k)\Big/\ln\frac{1}{\lambda(\mathbf{H})} = t$. As $t \geq \max\{|\frac{1+\ln\ln\frac{1}{\lambda(\mathbf{H})}-\ln 2MC_{\mathbf{H}}B^2}{\ln\frac{1}{\lambda(\mathbf{H})}}|, e^{|\ln 2MC_{\mathbf{H}}B^2|}/4, 16\}$, Lemma 5 gives us $\sharp(S_t) \leq k(t)-t \leq 3\frac{\ln t}{\ln(1/\lambda(\mathbf{H}))}$, where $\sharp(S_t)$ denotes the cardinality of set S_t. Using Equation (53) with $a=\frac{e}{\ln\frac{1}{\lambda(\mathbf{H})}}$ and $t=k$, we get

$$\ln k \leq \frac{k\ln\frac{1}{\lambda(\mathbf{H})}}{e} - \ln\ln\frac{1}{\lambda(\mathbf{H})} \Rightarrow \frac{\ln k}{\ln\frac{1}{\lambda(\mathbf{H})}} \leq \frac{k}{e} - \frac{\ln\ln\frac{1}{\lambda(\mathbf{H})}}{\ln\frac{1}{\lambda(\mathbf{H})}},$$

based on which we have $\frac{\ln 2MC_{\mathbf{H}}B^2k}{\ln\frac{1}{\lambda(\mathbf{H})}} = \frac{\ln 2MC_{\mathbf{H}}B^2}{\ln\frac{1}{\lambda(\mathbf{H})}} + \frac{\ln k}{\ln\frac{1}{\lambda(\mathbf{H})}} \leq \frac{k}{e} - \frac{\ln\ln\frac{1}{\lambda(\mathbf{H})}}{\ln\frac{1}{\lambda(\mathbf{H})}} + \frac{\ln 2MC_{\mathbf{H}}B^2}{\frac{1}{\lambda(\mathbf{H})}}$. If

$$k \geq \frac{2e\Big(\Big|\frac{\ln\ln\frac{1}{\lambda(\mathbf{H})}}{\ln\frac{1}{\lambda(\mathbf{H})}}\Big|+\Big|\frac{\ln 2MC_{\mathbf{H}}B^2}{\frac{1}{\lambda(\mathbf{H})}}\Big|\Big)}{e-2}, \tag{36}$$

we have $\frac{\ln 2MC_{\mathbf{H}}B^2k}{\ln\frac{1}{\lambda(\mathbf{H})}} \leq \frac{k}{e} + \Big|\frac{\ln\ln\frac{1}{\lambda(\mathbf{H})}}{\ln\frac{1}{\lambda(\mathbf{H})}}\Big| + \Big|\frac{\ln 2MC_{\mathbf{H}}B^2}{\frac{1}{\lambda(\mathbf{H})}}\Big| \leq \frac{k}{e}+\frac{(e-2)k}{2e} = \frac{k}{2}$, which indicates $\mathcal{T}_k \leq \frac{k}{2}$. Recall the definition $S_t := \{k\in\mathbb{Z}^+ \mid k-\mathcal{T}_k \leq t \leq k-1,\ k\geq K_0\}$, we then have $k \leq 2t,\ \ \forall k \in S_t$. Combining Equations (21), (25), (35) and (36), we consider

$$\begin{aligned}
k \geq \tau(\mathbf{H}) := \max\Bigg\{&\frac{[\frac{1}{\lambda(\mathbf{H})}]^{K_{\mathbf{H}}}}{2MC_{\mathbf{H}}B^2}, K_{\mathbf{H}}, \frac{2\ln\frac{e\ln\frac{1}{\lambda(\mathbf{H})}}{4MC_{\mathbf{H}}B^2}}{\ln\frac{1}{\lambda(\mathbf{H})}},\\
&\frac{2e\Big(\Big|\frac{\ln\ln\frac{1}{\lambda(\mathbf{H})}}{\ln\frac{1}{\lambda(\mathbf{H})}}\Big|+\Big|\frac{\ln 2MC_{\mathbf{H}}B^2}{\frac{1}{\lambda(\mathbf{H})}}\Big|\Big)}{e-2}, \frac{2\ln\frac{4MC_{\mathbf{H}}B^2}{e\ln\frac{1}{\lambda(\mathbf{H})}}}{\ln\frac{1}{\lambda(\mathbf{H})}}, 16,\\
&\Big|\frac{1+\ln\ln\frac{1}{\lambda(\mathbf{H})}-\ln 2MC_{\mathbf{H}}B^2}{\ln\frac{1}{\lambda(\mathbf{H})}}\Big|, e^{|\ln 2MC_{\mathbf{H}}B^2|}/4\Bigg\}.
\end{aligned} \tag{37}$$

Note that $\mathcal{T}_k$ increases with respect to k, we then have $\mathcal{T}_k \leq \mathcal{T}_{2t},\ \ \forall k\in S_t$. That means in $\sum_{k=\tau(\mathbf{H})}^{+\infty}(\mathcal{T}_k\sum_{d=k-\mathcal{T}_k}^{k-1}\gamma_d^2)$, γ_t^2 appears at most $\mathcal{T}_{2t}\cdot\sharp(S_t) \leq 3\ln(4MC_{\mathbf{H}}B^2t)\frac{\ln t}{\ln^2(1/\lambda(\mathbf{H}))}$. The direct calculations then give us

$$\sum_{k=\tau(\mathbf{H})}^{+\infty}\left(\mathcal{T}_k\sum_{d=k-\mathcal{T}_k}^{k-1}\gamma_d^2\right) \leq \sum_{t=\tau(\mathbf{H})}^{+\infty}\frac{3\ln(4MC_{\mathbf{H}}B^2t)\ln t\cdot\gamma_t^2}{\ln^2(1/\lambda(\mathbf{H}))}. \tag{38}$$

Going back to Equation (34) and Lemma 3, we are then led to

$$\begin{aligned}
&\sum_{k=\tau(\mathbf{H})}^{+\infty}\gamma_k\mathbb{E}\|\nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}})\|^2 \leq f(\overline{\mathbf{x}^0})-\min f\\
&+\Big(2B/\sqrt{m}+2B^2/m+2B^2LC_{\mathbf{W}}\Big)\\
&\quad\cdot\sum_{t=\tau(\mathbf{H})}^{+\infty}3\ln(4MC_{\mathbf{H}}B^2t)\frac{\ln t\cdot\gamma_t^2}{\ln^2(1/\lambda(\mathbf{H}))}.
\end{aligned} \tag{39}$$

Furthermore, with the boundedness $\|\nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}})\|^2 \leq B^2$, we can get

$$\begin{aligned}
&\sum_{k=1}^{+\infty}\gamma_k\mathbb{E}\|\nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}})\|^2 = \sum_{k=1}^{\tau(\mathbf{H})-1}\gamma_k\mathbb{E}\|\nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}})\|^2\\
&+\sum_{k=\tau(\mathbf{H})}^{+\infty}\gamma_k\mathbb{E}\|\nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}})\|^2 \leq \sum_{t=1}^{\tau(\mathbf{H})-1}\gamma_tB^2\\
&+\Big(2B/\sqrt{m}+2B^2/m+\frac{2B^2LC_{\mathbf{W}}}{\sqrt{m}}\Big)\\
&\quad\cdot\sum_{t=\tau(\mathbf{H})}^{+\infty}3\ln(4MC_{\mathbf{H}}B^2t)\frac{\ln t\cdot\gamma_t^2}{\ln^2(1/\lambda(\mathbf{H}))}+f(\overline{\mathbf{x}^0})-\min f\\
&=\mathcal{O}\Big(1+\frac{1}{\ln^2(1/\lambda(\mathbf{H}))}\cdot[1+\frac{1}{\sqrt{m}(1-\lambda(\mathbf{W}))}]\\
&\quad+f(\overline{\mathbf{x}^0})-\min f\Big).
\end{aligned} \tag{40}$$

Step 4: With Lipschitz of ∇f, it holds that

$$\begin{aligned}
&\gamma_k\|\nabla f(\overline{\mathbf{x}^k})\|^2-\gamma_k\|\nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}})\|^2\\
&=\gamma_k\langle\nabla f(\overline{\mathbf{x}^k})-\nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}}), \nabla f(\overline{\mathbf{x}^k})+\nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}})\rangle\\
&\leq\gamma_k\|\nabla f(\overline{\mathbf{x}^k})-\nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}})\|\cdot\|\nabla f(\overline{\mathbf{x}^k})+\nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}})\|\\
&\leq 2BL\gamma_k\|\overline{\mathbf{x}^k}-\overline{\mathbf{x}^{k-\mathcal{T}_k}}\| \overset{a)}{\leq} BL\gamma_k^2+BL\|\overline{\mathbf{x}^k}-\overline{\mathbf{x}^{k-\mathcal{T}_k}}\|^2,
\end{aligned}$$

where a) uses Cauchy's inequality $2ab \leq a^2 + b^2$ with $a = \sqrt{BL}\gamma_k$ and $b = \sqrt{BL}\|\overline{\mathbf{x}^k} - \overline{\mathbf{x}^{k-\mathcal{T}_k}}\|$. Taking the expectation of both sides of the inequality above gives us

$$\begin{aligned}&\gamma_k\mathbb{E}\|\nabla f(\overline{\mathbf{x}^k})\|^2 - \gamma_k\mathbb{E}\|\nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}})\|^2\\ &\leq BL\gamma_k^2 + BL\mathbb{E}\|\overline{\mathbf{x}^k} - \overline{\mathbf{x}^{k-\mathcal{T}_k}}\|^2.\end{aligned} \tag{41}$$

Direct computations give us

$$\begin{aligned}&\sum_{k=1}^{+\infty}\mathbb{E}\|\overline{\mathbf{x}^k} - \overline{\mathbf{x}^{k-\mathcal{T}_k}}\|^2 \leq \mathcal{T}_k\sum_{k=1}^{+\infty}\sum_{d=k-\mathcal{T}_k}^{k}\mathbb{E}\|\overline{\mathbf{x}^{d+1}} - \overline{\mathbf{x}^d}\|^2\\
&\leq B^2/m\sum_{k=1}^{+\infty}\mathcal{T}_k\sum_{d=k-\mathcal{T}_k}^{k}\gamma_d^2 = B^2/m\sum_{k=1}^{\tau(\mathbf{H})-1}\mathcal{T}_k\sum_{d=k-\mathcal{T}_k}^{k}\gamma_d^2\\
&\quad + B^2/m\sum_{k=\tau(\mathbf{H})}^{\infty}\mathcal{T}_k\sum_{d=k-\mathcal{T}_k}^{k}\gamma_d^2\\
&\leq B^2[\tau(\mathbf{H})]^4/(4m) + B^2/m\sum_{k=\tau(\mathbf{H})}^{\infty}\mathcal{T}_k\sum_{d=k-\mathcal{T}_k}^{k}\gamma_d^2\\
&\overset{\text{Equation (38)}}{\leq} B^2[\tau(\mathbf{H})]^4/(4m)\\
&\quad + B^2/m\sum_{t=\tau(\mathbf{H})}^{+\infty}3\ln(4MC_{\mathbf{H}}B^2t)\frac{\ln t\cdot\gamma_t^2}{\ln^2(1/\lambda(\mathbf{H}))}.\end{aligned} \tag{42}$$

Thus, we can get

$$\begin{aligned}&\sum_{k=1}^{K}\gamma_k\mathbb{E}\|\nabla f(\overline{\mathbf{x}^k})\|^2 \leq \sum_{k=1}^{\infty}\gamma_k\mathbb{E}\|\nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}})\|^2\\
&\quad + \sum_{k=1}^{\infty}\big(\gamma_k\mathbb{E}\|\nabla f(\overline{\mathbf{x}^k})\|^2 - \gamma_k\mathbb{E}\|\nabla f(\overline{\mathbf{x}^{k-\mathcal{T}_k}})\|^2\big)\\
&\leq \sum_{t=1}^{\tau(\mathbf{H})-1}\gamma_tB^2 + B^3L[\tau(\mathbf{H})]^4/(4m) + BL\sum_{t=1}^{\infty}\gamma_t^2 + \sum_{t=1}^{\infty}\frac{\gamma_t}{2t}\\
&+ \Big(2B/\sqrt{m} + B^3L/m + 2B^2/m + \frac{2B^2LC_{\mathbf{W}}}{\sqrt{m}}\Big)\\
&\cdot\sum_{t=\tau(\mathbf{H})}^{+\infty}3\ln(4MC_{\mathbf{H}}B^2t)\frac{\ln t\cdot\gamma_t^2}{\ln^2(1/\lambda(\mathbf{H}))} + f(\overline{\mathbf{x}^0}) - \min f =\\
&\mathcal{O}\Big(1 + \frac{1}{\ln^2(1/\lambda(\mathbf{H}))}\cdot[\frac{1}{\sqrt{m}(1-\lambda(\mathbf{W}))}] + f(\overline{\mathbf{x}^0}) - \min f\Big)\\
&< +\infty.\end{aligned}$$

We bound $\sum_{k=1}^{K}\gamma_k$ as

$$\begin{aligned}&\sum_{k=1}^{K}\gamma_k \geq \sum_{k=1}^{K}\int_k^{k+1}\frac{1}{(t+1)^\theta}dt = \int_1^{K+1}\frac{1}{(t+1)^\theta}dt\\
&= \frac{(K+1)^{1-\theta} - 2^{1-\theta}}{1-\theta} \geq (1 - \frac{1}{2^{1-\theta}})(K+1)^{1-\theta},\end{aligned}$$

as $K \geq 3$. Thus, the proof of the convergence rate is completed by $1/(\sum_{k=1}^{K}\gamma_k) \leq \frac{1}{1-\frac{1}{2^{1-\theta}}}\cdot\frac{1}{(K+1)^{1-\theta}}$, as $K \geq 3$. With the fact $\min_{1\leq k\leq K}\{\mathbb{E}\|\nabla f(\overline{\mathbf{x}^k})\|^2\} \leq \frac{\sum_{k=1}^{K}\gamma_k\mathbb{E}\|\nabla f(\overline{\mathbf{x}^k})\|^2}{\sum_{k=1}^{K}\gamma_k}$, we then get the ergodic convergence rate.

Proof of the nonergodic convergence: Notice that $\overline{\mathbf{x}^{k+1}} - \overline{\mathbf{x}^k} = -\gamma_k\frac{\sum_{i=1}^{m}\nabla f_i^{j_{i,k}}(\mathbf{x}^k(i))}{m}$, we have

$$\begin{aligned}\|\overline{\mathbf{x}^{k+1}} - \overline{\mathbf{x}^k}\| &= \gamma_k\Big\|\frac{\sum_{i=1}^{m}\nabla f_i^{j_{i,k}}(\mathbf{x}^k(i))}{m}\Big\|\\
&\leq \frac{\gamma_k}{m}\sum_{i=1}^{m}\|\nabla f_i^{j_{i,k}}(\mathbf{x}^k(i))\| \leq B\cdot\gamma_k.\end{aligned} \tag{43}$$

Based on Equation (43), we can get

$$\begin{aligned}&\Big|\|\nabla f(\overline{\mathbf{x}^{k+1}})\|^2 - \|\nabla f(\overline{\mathbf{x}^k})\|^2\Big|\\
&= \Big|\langle\nabla f(\overline{\mathbf{x}^{k+1}}) - \nabla f(\overline{\mathbf{x}^k}), \nabla f(\overline{\mathbf{x}^{k+1}}) + \nabla f(\overline{\mathbf{x}^k})\rangle\Big|\\
&\leq \|\nabla f(\overline{\mathbf{x}^{k+1}}) + \nabla f(\overline{\mathbf{x}^k})\|\cdot\|\nabla f(\overline{\mathbf{x}^{k+1}}) - \nabla f(\overline{\mathbf{x}^k})\|\\
&\leq 2BL\|\overline{\mathbf{x}^{k+1}} - \overline{\mathbf{x}^k}\| = 2B^2L\gamma_k.\end{aligned} \tag{44}$$

Thus, we derive

$$\begin{aligned}&\Big|\mathbb{E}\|\nabla f(\overline{\mathbf{x}^{k+1}})\|^2 - \mathbb{E}\|\nabla f(\overline{\mathbf{x}^k})\|^2\Big|\\
&\leq \mathbb{E}\Big|\|\nabla f(\overline{\mathbf{x}^{k+1}})\|^2 - \|\nabla f(\overline{\mathbf{x}^k})\|^2\Big| = 2B^2L\gamma_k.\end{aligned} \tag{45}$$

Notice that we have proved $\sum_k\gamma_k\mathbb{E}\|\nabla f(\overline{\mathbf{x}^k})\|^2 < +\infty$, and Equation (45) has shown $\Big|\mathbb{E}\|\nabla f(\overline{\mathbf{x}^{k+1}})\|^2 - \mathbb{E}\|\nabla f(\overline{\mathbf{x}^k})\|^2\Big| = 2B^2L\gamma_k$. Recalling Lemma 4 with $h_k \leftarrow \gamma_k$ and $\beta_k \leftarrow \mathbb{E}\|\nabla f(\overline{\mathbf{x}^k})\|^2$, we then get $\lim_k\mathbb{E}\|\nabla f(\overline{\mathbf{x}^k})\| = 0$.

C. Proof of Proposition 1

First, it is easy to see $C_{\mathbf{W}} = \mathcal{O}(\frac{\sqrt{\ln(1/\lambda(\mathbf{H}))}}{1-\lambda(\mathbf{W})})$ in Lemma 3 with the stepsizes. Applying the stepsizes choice Equation (13) to the upper Equation (42) directly gives us

$$\begin{aligned}&\sum_{k=1}^{\infty}\gamma_k\mathbb{E}\|\nabla f(\overline{\mathbf{x}^k})\|^2 \leq \sum_{t=1}^{\tau(\mathbf{H})-1}\gamma_tB^2 + B^3L[\tau(\mathbf{H})]^4/4\\
&+ BL\sum_{t=1}^{\infty}\gamma_t^2 + \Big(2B + BL + 2B^2 + \frac{2B^2LC_{\mathbf{W}}}{\sqrt{m}}\Big)\\
&\cdot\sum_{t=\tau(\mathbf{H})}^{+\infty}3\ln(4MC_{\mathbf{H}}B^2t)\frac{\ln t\cdot\gamma_t^2}{\ln^2(1/\lambda(\mathbf{H}))} + f(\overline{\mathbf{x}^0}) - \min f\\
&= \mathcal{O}\Big(1 + \frac{1}{\sqrt{m}(1-\lambda(\mathbf{W}))\sqrt{\ln(1/\lambda(\mathbf{H}))}} + f(\overline{\mathbf{x}^0}) - \min f\Big).\end{aligned}$$

Further with the fact $\sum_{i=1}^{k}\gamma_i = \Theta[\sqrt{\ln(1/\lambda(\mathbf{H}))}k^{1-\theta}]$, we then get the result.

D. Proof of Proposition 2

The proof of Proposition 2 is similar to the proof of Theorem 1. Note that

$$\frac{1}{m}\sum_{i=1}^{m}\|\mathbf{x}^k(i) - \overline{\mathbf{x}^k}\| = \mathcal{O}(\frac{1}{1-\lambda(\mathbf{W})}\cdot\frac{1}{(k+1)^\theta})$$

still holds for scheme Equation (2), if the stepsizes are selected as Equation (9) and Assumption 5 holds.

Like previous proofs, the proof also consists of four steps: 1. Introduce the delay $\mathcal{H}_k$. 2. the second one is to Bound

$\sum_k \gamma_k \mathbb{E}\|\nabla \mathcal{F}(\overline{\mathbf{x}^{k-\mathcal{T}_k}})\|^2$ by four terms. 3. Prove the summability of the four terms. 4. Establish an upper bound for $\sum_k \left(\gamma_k \mathbb{E}\|\nabla \mathcal{F}(\overline{\mathbf{x}^k})\|^2 - \gamma_k \mathbb{E}\|\nabla \mathcal{F}(\overline{\mathbf{x}^{k-\mathcal{T}_k}})\|^2\right)$.

Step 1. Assume that C^i and λ_i are the factors in Equation (14) for Markov chain in the i-th node. Let $C := \max\{C^1, C^2, \ldots, C^m\}$ and $\lambda := \max\{\lambda_1, \lambda_2, \ldots, \lambda_m\}$. For integer $k \geq 1$, we consider the integer $\mathcal{H}_k$ as

$$\mathcal{H}_k := \min\{\left\lceil \ln(2CB^2k)/\ln(1/\lambda)\right\rceil, k\}. \tag{46}$$

It is easy to see $\mathcal{H}_k \leq k$. With [Theorem 4.9, [17]], we have the following inequality

$$\int_\Xi |p_s^{s+\mathcal{H}_k}(\xi) - \pi(\xi)| d\mu(\xi) \leq \frac{1}{2 \cdot B^2 \cdot k}, \forall s \in \mathbb{Z}^+,$$

where $p_s^{s+\mathcal{H}_k}(\xi)$ denotes the transition probability density function (p.d.f.) from s to $s+\mathcal{H}_k$ with respect to ξ. The property of time-homogeneous of the Markov chain directly gives that $p_s^{s+\mathcal{H}_k}(\xi) = p_0^{\mathcal{H}_k}(\xi)$, i.e.,

$$\int_\Xi |p_0^{\mathcal{H}_k}(\xi) - \pi(\xi)| d\mu(\xi) \leq \frac{1}{2 \cdot B^2 \cdot k}, \forall s \in \mathbb{Z}^+ \tag{47}$$

Step 2. Denote the shorthand notation $\tilde{\mathbf{d}}^{k-\mathcal{H}_k} := \frac{1}{m}\sum_{i=1}^m \nabla F(\mathbf{x}^{k-\mathcal{H}_k}(i); \xi^k(i))$, we calculate the lower bound for following inner product:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{j}_k}(\langle \nabla \mathcal{F}(\overline{\mathbf{x}^{k-\mathcal{H}_k}}), \tilde{\mathbf{d}}^{k-\mathcal{H}_k}\rangle \mid \chi^{k-\mathcal{H}_k}) \\
&\overset{a)}{=} \left\langle \nabla \mathcal{F}(\overline{\mathbf{x}^{k-\mathcal{H}_k}}), \frac{1}{m}\sum_{i=1}^m \nabla F(\mathbf{x}^{k-\mathcal{H}_k}(i); \xi) p_{k-\mathcal{H}_k}^k(\xi) d\mu(\xi) \right\rangle \\
&\overset{b)}{=} \left\langle \nabla \mathcal{F}(\overline{\mathbf{x}^{k-\mathcal{H}_k}}), \frac{1}{m}\sum_{i=1}^m \nabla F(\mathbf{x}^{k-\mathcal{H}_k}(i); \xi) p_0^{\mathcal{H}_k}(\xi) d\mu(\xi) \right\rangle \\
&= \left\langle \nabla \mathcal{F}(\overline{\mathbf{x}^{k-\mathcal{H}_k}}), \frac{1}{m}\sum_{i=1}^m \nabla F(\mathbf{x}^{k-\mathcal{H}_k}(i); \xi) \pi(\xi) d\mu(\xi) \right\rangle \\
&+ \left\langle \nabla \mathcal{F}(\overline{\mathbf{x}^{k-\mathcal{H}_k}}), \frac{1}{m}\sum_{i=1}^m \nabla F(\mathbf{x}^{k-\mathcal{H}_k}(i); \xi) \right. \\
&\qquad \left. \cdot [p_0^{\mathcal{H}_k}(\xi) - \pi(\xi)] d\mu(\xi) \right\rangle \\
&\overset{c)}{\geq} \|\nabla \mathcal{F}(\overline{\mathbf{x}^{k-\mathcal{H}_k}})\|^2 - \frac{1}{2k} + \left\langle \nabla \mathcal{F}(\overline{\mathbf{x}^{k-\mathcal{H}_k}}), \right. \\
&\qquad \left. \frac{1}{m}\sum_{i=1}^m [\nabla F(\mathbf{x}^{k-\mathcal{H}_k}(i); \xi) - \nabla F(\overline{\mathbf{x}^{k-\mathcal{H}_k}}; \xi)] \pi(\xi) d\mu(\xi) \right\rangle \\
&\overset{d)}{\geq} \|\nabla \mathcal{F}(\overline{\mathbf{x}^{k-\mathcal{H}_k}})\|^2 - \frac{1}{2k} - \frac{BL}{m}\sum_{i=1}^m \|\mathbf{x}^{k-\mathcal{H}_k}(i) - \overline{\mathbf{x}^{k-\mathcal{H}_k}}\|,
\end{aligned} \tag{48}$$

where $a)$ uses the conditional expectation, and $b)$ comes from the property of Markov chain, and $c)$ depends on Equation (47), and $d)$ is due to the Lipschitz property. Rearranging Equation (48) gives us

$$\begin{aligned}
\gamma_k \mathbb{E}\|\nabla \mathcal{F}(\overline{\mathbf{x}^{k-\mathcal{H}_k}})\|^2 &\leq \gamma_k \mathbb{E}(\langle \nabla \mathcal{F}(\overline{\mathbf{x}^{k-\mathcal{H}_k}}), \tilde{\mathbf{d}}^{k-\mathcal{H}_k}\rangle) \\
&+ \frac{\gamma_k}{2k} + \gamma_k \frac{BL}{m}\sum_{i=1}^m \mathbb{E}\|\mathbf{x}^{k-\mathcal{H}_k}(i) - \overline{\mathbf{x}^{k-\mathcal{H}_k}}\|.
\end{aligned} \tag{49}$$

Then we need to bound $\sum_k \gamma_k \mathbb{E}(\langle \nabla \mathcal{F}(\overline{\mathbf{x}^{k-\mathcal{H}_k}}), \tilde{\mathbf{d}}^{k-\mathcal{H}_k}\rangle)$, the rest part is almost identical to the one of previous proof and will not be repeated.

Step 3 and Step 4. These two parts are very similar to ones of Theorem 1 and will not be repeated.

VII. Concluding Remarks

In this paper, we propose decentralized Markov chain gradient descent, where the samples are drawn along a trajectory of a Markov chain over the network. Our proposed algorithms can be used when it is expensive or even impossible to sample directly from a distribution, but sampling via a Markov chain is possible and relatively cheap. The convergence analysis is established in various settings.

Appendix

We define $\mathbf{1} := [1, 1, \cdots, 1]^\top \in \mathbb{R}^m$, and the projection matrix is given as $\mathbf{P} := \frac{\mathbf{1}\mathbf{1}^\top}{m} \in \mathbb{R}^{m\times m}$. It is easy to see

$$\mathbf{W}\mathbf{P} = \mathbf{P}\mathbf{W} = \mathbf{P}. \tag{50}$$

A. Proof of Lemma 3

The global scheme can be described as $\mathbf{x}^{k+1} = \mathbf{W}\mathbf{x}^k - \gamma_k \mathbf{u}^k$, where $\mathbf{u}^k$ has been defined in Equation (4).

With direct calculation, we have

$$\mathbf{x}^k = \mathbf{W}^k \mathbf{x}^0 - \sum_{j=0}^{k-1} \gamma_j \mathbf{W}^{k-j} \mathbf{u}^j. \tag{51}$$

Recalling Equation (50), i.e., $\mathbf{P}\mathbf{W} = \mathbf{P}$, we have

$$\mathbf{P}\mathbf{W}^{k-j} = \mathbf{P}\mathbf{W}\mathbf{W}^{k-j-1} = \mathbf{P}\mathbf{W}^{k-j-1} = \ldots = \mathbf{P}$$

as $j < k$. Further using the direct expansion of Equation (51) and the fact that initializations are all selected from $\mathbf{B}(\mathbf{0}, B)$, we have

$$\begin{aligned}
\|(\mathbf{I}-\mathbf{P})\mathbf{x}^k\| &= \|(\mathbf{W}^k - \mathbf{P})\mathbf{x}^0 - \sum_{j=0}^{k-1} \gamma_j (\mathbf{W}^{k-j} - \mathbf{P})\mathbf{u}^j\| \\
&\leq \|\mathbf{W}^k - \mathbf{P}\| \cdot \|\mathbf{x}^0\| + \sum_{j=0}^{k-1} \gamma_j \|\mathbf{W}^{k-j} - \mathbf{P}\| \cdot \|\mathbf{u}^j\| \\
&\leq \max_k \left\{ \sqrt{\sum_{i=1}^m \|\nabla f_i^{j_{i,k}}(\mathbf{x}^k(i))\|^2}, \sqrt{m}B \right\} \cdot \sum_{j=0}^k \gamma_j \lambda(\mathbf{W})^{k-j} \\
&\leq \sqrt{m}B \sum_{j=0}^k \gamma_j \lambda(\mathbf{W})^{k-j},
\end{aligned}$$

where the second inequality uses Lemma 2 as $\|\mathbf{W}^{k-j} - \mathbf{P}\| \leq \lambda(\mathbf{W})^{k-j}$. Notice that $\|(\mathbf{I}-\mathbf{P})\mathbf{x}^k\|^2 = \sum_{i=1}^m \|\mathbf{x}^k(i) - \overline{\mathbf{x}^k}\|^2$, we have

$$\begin{aligned}
\frac{1}{m}\sum_{i=1}^m \|\mathbf{x}^k(i) - \overline{\mathbf{x}^k}\| &\leq \frac{\sqrt{m\sum_{i=1}^m \|\mathbf{x}^k(i) - \overline{\mathbf{x}^k}\|^2}}{m} \\
&\leq \frac{1}{\sqrt{m}}\|(\mathbf{I}-\mathbf{P})\mathbf{x}^k\| \leq B\sum_{j=0}^k \gamma_j \lambda(\mathbf{W})^{k-j}.
\end{aligned}$$

If $\gamma_j = \frac{1}{(j+1)^\theta}$, it holds $\gamma_j \leq \frac{2^\theta}{(k+1)^\theta}$ as $j \geq \lceil \frac{k}{2} \rceil + 1 \geq \frac{k-1}{2}$, and we have

$$\begin{aligned}
&\sum_{j=0}^{k} \gamma_j \lambda(\mathbf{W})^{k-j} = \sum_{j=0}^{\lceil \frac{k}{2} \rceil} \gamma_j \lambda(\mathbf{W})^{k-j} + \sum_{j=\lceil \frac{k}{2} \rceil + 1}^{k} \gamma_j \lambda(\mathbf{W})^{k-j} \\
&\leq k\lambda(\mathbf{W})^{k/2} + \frac{2^\theta}{(k+1)^\theta} \sum_{j=\lceil \frac{k}{2} \rceil + 1}^{k} \lambda(\mathbf{W})^{k-j} \\
&\leq k\lambda(\mathbf{W})^{k/2} + \frac{2^\theta}{(k+1)^\theta} \frac{1}{1-\lambda(\mathbf{W})} \leq \frac{C_{\mathbf{W}}}{(k+1)^\theta},
\end{aligned}$$

where

$$C_{\mathbf{W}} := \frac{2^\theta}{1-\lambda(\mathbf{W})} + \sup_k \{(k+1)^{1+\theta} \lambda(\mathbf{W})^{k/2}\}. \tag{52}$$

Now, we turn to show the finiteness of $\sup_k\{(k+1)^{1+\theta}\lambda(\mathbf{W})^{k/2}\}$. By setting $t = k+1$, with Lemma 6, it follows that $k+1 \leq \frac{a}{e} e^{\frac{k+1}{a}}$ with $a > 0$ to be determined. By setting $a = \frac{2+2\theta}{\ln \frac{1}{\lambda(\mathbf{W})}} > 0$, with the fact $0 \leq \lambda(\mathbf{W}) < 1$, it holds $0 < a < \frac{2+2\theta}{\ln 2}$. Then for any $k \geq 1$, we are then led to

$$\begin{aligned}
&(k+1)^{1+\theta}\lambda(\mathbf{W})^{k/2} \leq (\frac{a}{e})^{1+\theta} e^{\frac{1+\theta}{a}(k+1)} \lambda(\mathbf{W})^{k/2} \\
&= a^{1+\theta} e^{(1-\frac{1}{a})(1+\theta)} e^{\frac{1+\theta}{a}k} e^{(k \ln \lambda(\mathbf{W}))/2} \\
&= a^{1+\theta} e^{(1-\frac{1}{a})(1+\theta)} e^{(\frac{1+\theta}{a} + \frac{1}{2}\ln \lambda(\mathbf{W}))k} \\
&= a^{1+\theta} e^{(1-\frac{1}{a})(1+\theta)} \leq (\frac{2+2\theta}{\ln 2})^{1+\theta} e^{1+\theta}.
\end{aligned}$$

Thus, we get $C_{\mathbf{W}} = \mathcal{O}(\frac{1}{1-\lambda(\mathbf{W})})$.

B. Proof of Lemma 5

As $y \geq e^{\frac{|c|}{a}}$, it holds that

$$a \ln y - c \leq |a \ln y - c| \leq a \ln y + |c| \leq 2a \ln y.$$

In Equation (53) with $a \leftarrow 2a$ and $t \leftarrow y$, $\ln y \leq y/2a + \ln 2a - 1 \leq y/2$. Considering the fact $y - a \ln y + c = x$, we have

$$\begin{aligned}
&y + c = x + a \ln y \leq x + y/2 + a(\ln 2a - 1) \\
&\Rightarrow y/2 \leq x + a(\ln 2a - 1) - c.
\end{aligned}$$

Thus, if $x \geq |a(\ln 2a - 1) - c|$, we then get $y \leq 4x$. In summary, as $x \geq \max\{|a(\ln 2a - 1) - c|, e^{\frac{|c|}{a}}/4, 16\}$,

$$y - x = a \ln y - c \leq 2a \ln y = 2a \ln x + 4a \ln 2 \leq 3a \ln x.$$

C. Proof of Lemma 6

Consider the function $g(t) := \ln t - \frac{t}{a} - \ln a + 1$. Noticing that $g'(t) = \frac{1}{t} - \frac{1}{a}$, $g(t)$ gets the maximum at $t = a$. We then get

$$\ln t \leq \frac{t}{a} + \ln a - 1, \tag{53}$$

which also yields

$$t = e^{\ln t} \leq e^{\frac{t}{a} + \ln a - 1} = \frac{a}{e} e^{\frac{t}{a}}.$$

REFERENCES

[1] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, A. Senior, P. Tucker, K. Yang, Q. V. Le *et al.*, "Large scale distributed deep networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1223–1231.

[2] M. Li, D. G. Andersen, J. W. Park, A. J. Smola, A. Ahmed, V. Josifovski, J. Long, E. J. Shekita, and B.-Y. Su, "Scaling distributed machine learning with the parameter server." in *OSDI*, vol. 14, 2014, pp. 583–598.

[3] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification." *Journal of machine learning research*, vol. 10, no. 2, 2009.

[4] B. Kulis *et al.*, "Metric learning: A survey," *Foundations and Trends® in Machine Learning*, vol. 5, no. 4, pp. 287–364, 2013.

[5] E. Xing, M. Jordan, S. J. Russell, and A. Ng, "Distance metric learning with application to clustering with side-information," *Advances in neural information processing systems*, vol. 15, 2002.

[6] Y. Ying and P. Li, "Distance metric learning with eigenvalue optimization," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 1–26, 2012.

[7] P. Zhao, S. C. Hoi, R. Jin, and T. Yang, "Online AUC maximization," in *Proceedings of the 28th International Conference on Machine Learning ICML*, 2011, pp. 233–240.

[8] W. Gao, R. Jin, S. Zhu, and Z.-H. Zhou, "One-pass AUC optimization," in *International conference on machine learning*. PMLR, 2013, pp. 906–914.

[9] Y. Ying, L. Wen, and S. Lyu, "Stochastic online AUC maximization," *Advances in neural information processing systems*, vol. 29, 2016.

[10] M. Liu, X. Zhang, Z. Chen, X. Wang, and T. Yang, "Fast stochastic AUC maximization with $O(1/n)$-convergence rate," in *International Conference on Machine Learning*. PMLR, 2018, pp. 3189–3197.

[11] W. Rejchel, "On ranking and generalization bounds." *Journal of Machine Learning Research*, vol. 13, no. 5, 2012.

[12] S. Agarwal and P. Niyogi, "Generalization bounds for ranking algorithms via algorithmic stability." *Journal of Machine Learning Research*, vol. 10, no. 2, 2009.

[13] Z. Yang, Y. Lei, P. Wang, T. Yang, and Y. Ying, "Simple stochastic and online gradient descent algorithms for pairwise learning," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[14] B. Poljak and J. Z. Tsypkin, "Robust identification," *Automatica*, vol. 16, no. 1, pp. 53–63, 1980.

[15] T. Sarkar and A. Rakhlin, "Near optimal finite time identification of arbitrary linear dynamical systems," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5610–5618.

[16] X. Lian, C. Zhang, H. Zhang, C.-J. Hsieh, W. Zhang, and J. Liu, "Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent," in *Advances in Neural Information Processing Systems*, 2017, pp. 5330–5340.

[17] R. Montenegro, P. Tetali *et al.*, "Mathematical aspects of mixing times in markov chains," *Foundations and Trends® in Theoretical Computer Science*, vol. 1, no. 3, pp. 237–354, 2006.

[18] T. Sun, Y. Sun, and W. Yin, "On markov chain gradient descent," in *Advances in Neural Information Processing Systems*, 2018, pp. 9917–9926.

[19] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Gossip algorithms: Design, analysis and applications," in *INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings IEEE*, vol. 3. IEEE, 2005, pp. 1653–1664.

[20] R. Olfati-Saber, J. A. Fax, and R. M. Murray, "Consensus and cooperation in networked multi-agent systems," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 215–233, 2007.

[21] L. Schenato and G. Gamba, "A distributed consensus protocol for clock synchronization in wireless sensor network," 2007.

[22] T. C. Aysal, M. E. Yildiz, A. D. Sarwate, and A. Scaglione, "Broadcast gossip algorithms for consensus," *IEEE Transactions on Signal processing*, vol. 57, no. 7, pp. 2748–2761, 2009.

[23] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.

[24] A. I. Chen and A. Ozdaglar, "A fast distributed proximal-gradient method," in *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*. IEEE, 2012, pp. 601–608.

[25] D. Jakovetić, J. Xavier, and J. M. Moura, "Fast distributed gradient methods," *IEEE Transactions on Automatic Control*, vol. 59, no. 5, pp. 1131–1146, 2014.

[26] I. Matei and J. S. Baras, "Performance evaluation of the consensus-based distributed subgradient method under random communication topologies," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 4, pp. 754–771, 2011.

[27] K. Yuan, Q. Ling, and W. Yin, "On the convergence of decentralized gradient descent," *SIAM Journal on Optimization*, vol. 26, no. 3, pp. 1835–1854, 2016.

[28] T.-H. Chang, M. Hong, and X. Wang, "Multi-agent distributed optimization via inexact consensus admm." *IEEE Trans. Signal Processing*, vol. 63, no. 2, pp. 482–497, 2015.

[29] I. D. Schizas, A. Ribeiro, and G. B. Giannakis, "Consensus in ad hoc wsns with noisy links part i: Distributed estimation of deterministic signals," *IEEE Transactions on Signal Processing*, vol. 56, no. 1, pp. 350–364, 2008.

[30] W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin, "On the linear convergence of the admm in decentralized consensus optimization." *IEEE Trans. Signal Processing*, vol. 62, no. 7, pp. 1750–1761, 2014.

[31] W. Shi, Q. Ling, G. Wu, and W. Yin, "Extra: An exact first-order algorithm for decentralized consensus optimization," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 944–966, 2015.

[32] J. Zeng and W. Yin, "On nonconvex decentralized gradient descent," *IEEE Transactions on Signal Processing*, vol. 66, no. 11, pp. 2834–2848, 2018.

[33] S. Hosseini, A. Chapman, and M. Mesbahi, "Online distributed convex optimization on dynamic networks." *IEEE Trans. Automat. Contr.*, vol. 61, no. 11, pp. 3545–3550, 2016.

[34] B. McMahan and M. Streeter, "Delay-tolerant algorithms for asynchronous distributed online learning," in *Advances in Neural Information Processing Systems*, 2014, pp. 2915–2923.

[35] T. Sun and D. Li, "Capri: Consensus accelerated proximal reweighted iteration for a class of nonconvex minimizations," *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[36] S. S. Ram, A. Nedić, and V. V. Veeravalli, "Asynchronous gossip algorithm for stochastic optimization: Constant stepsize analysis," in *Recent Advances in Optimization and its Applications in Engineering*. Springer, 2010, pp. 51–60.

[37] K. Srivastava and A. Nedic, "Distributed asynchronous constrained stochastic optimization," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 4, pp. 772–790, 2011.

[38] S. S. Ram, A. Nedić, and V. V. Veeravalli, "Distributed stochastic subgradient projection algorithms for convex optimization," *Journal of optimization theory and applications*, vol. 147, no. 3, pp. 516–545, 2010.

[39] B. Sirb and X. Ye, "Consensus optimization with delayed and stochastic gradients on decentralized networks," in *Big Data (Big Data), 2016 IEEE International Conference on*. IEEE, 2016, pp. 76–85.

[40] G. Lan, S. Lee, and Y. Zhou, "Communication-efficient algorithms for decentralized and stochastic optimization," *arXiv preprint arXiv:1701.03961*, 2017.

[41] X. Lian, W. Zhang, C. Zhang, and J. Liu, "Asynchronous decentralized parallel stochastic gradient descent," in *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 3043–3052.

[42] T. Sun, D. Li, and B. Wang, "Stability and generalization of decentralized stochastic gradient descent," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 11, 2021, pp. 9756–9764.

[43] ——, "Decentralized federated averaging," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[44] A. Koloskova, N. Loizou, S. Boreiri, M. Jaggi, and S. Stich, "A unified theory of decentralized sgd with changing topology and local updates," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5381–5393.

[45] A. Koloskova, S. Stich, and M. Jaggi, "Decentralized stochastic optimization and gossip algorithms with compressed communication," in *International Conference on Machine Learning*. PMLR, 2019, pp. 3478–3487.

[46] N. Singh, D. Data, J. George, and S. Diggavi, "Squarm-sgd: Communication-efficient momentum sgd for decentralized optimization," *IEEE Journal on Selected Areas in Information Theory*, vol. 2, no. 3, pp. 954–969, 2021.

[47] A. Koloskova, T. Lin, S. U. Stich, and M. Jaggi, "Decentralized deep learning with arbitrary communication compression," in *International Conference on Learning Representations*, 2019.

[48] X. Mao, K. Yuan, Y. Hu, Y. Gu, A. H. Sayed, and W. Yin, "Walkman: A communication-efficient random-walk algorithm for decentralized optimization," *IEEE Transactions on Signal Processing*, vol. 68, pp. 2513–2528, 2020.

[49] D. Kovalev, A. Koloskova, M. Jaggi, P. Richtarik, and S. Stich, "A linearly convergent algorithm for decentralized optimization: Sending less bits for free!" in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 4087–4095.

[50] B. Johansson, M. Rabi, and M. Johansson, "A randomized incremental subgradient method for distributed optimization in networked systems," *SIAM Journal on Optimization*, vol. 20, no. 3, pp. 1157–1170, 2009.

[51] ——, "A simple peer-to-peer algorithm for distributed optimization in sensor networks," in *Decision and Control, 2007 46th IEEE Conference on*. IEEE, 2007, pp. 4705–4710.

[52] S. S. Ram, A. Nedić, and V. V. Veeravalli, "Incremental stochastic subgradient algorithms for convex optimization," *SIAM Journal on Optimization*, vol. 20, no. 2, pp. 691–717, 2009.

[53] J. C. Duchi, A. Agarwal, M. Johansson, and M. I. Jordan, "Ergodic mirror descent," *SIAM Journal on Optimization*, vol. 22, no. 4, pp. 1549–1578, 2012.

[54] T. T. Doan, L. M. Nguyen, N. H. Pham, and J. Romberg, "Finite-time analysis of stochastic gradient descent under markov randomness," *arXiv preprint arXiv:2003.10973*, 2020.

[55] T. T. Doan, "Finite-time analysis of markov gradient descent," *IEEE Transactions on Automatic Control*, 2022.

[56] T. T. Doan, L. M. Nguyen, N. H. Pham, and J. Romberg, "Convergence rates of accelerated markov gradient descent with applications in reinforcement learning," *arXiv preprint arXiv:2002.02873*, 2020.

[57] T. Sun, D. Li, and B. Wang, "Adaptive random walk gradient descent for decentralized optimization," in *International Conference on Machine Learning*. PMLR, 2022, pp. 20 790–20 809.

[58] S. Boyd, P. Diaconis, and L. Xiao, "Fastest mixing markov chain on a graph," *SIAM review*, vol. 46, no. 4, pp. 667–689, 2004.

[59] C. B. Issaid, A. Elgabli, J. Park, M. Bennis, and M. Debbah, "Communication efficient decentralized learning over bipartite graphs," *IEEE Transactions on Wireless Communications*, vol. 21, no. 6, pp. 4150–4167, 2021.

[60] S. Pemmaraju, S. Skiena *et al.*, *Computational discrete mathematics: Combinatorics and graph theory with mathematica®*. Cambridge university press, 2003.

[61] A. Frieze and M. Karoński, *Introduction to random graphs*. Cambridge University Press, 2016.

[62] A. S. Asratian, T. M. Denley, and R. Häggkvist, *Bipartite graphs and their applications*. Cambridge university press, 1998, vol. 131.