A Sharp Estimate on the Transient Time of Distributed Stochastic Gradient Descent

Shi Pu, Alex Olshevsky, and Ioannis Ch. Paschalidis

***Abstract*—This paper is concerned with minimizing the average of n cost functions over a network in which agents may communicate and exchange information with each other. We consider the setting where only noisy gradient information is available. To solve the problem, we study the distributed stochastic gradient descent (DSGD) method and perform a non-asymptotic convergence analysis. For strongly convex and smooth objective functions, in expectation, DSGD asymptotically achieves the optimal network independent convergence rate compared to centralized stochastic gradient descent (SGD). Our main contribution is to characterize the transient time needed for DSGD to approach the asymptotic convergence rate. Moreover, we construct a "hard" optimization problem that proves the sharpness of the obtained result. Numerical experiments demonstrate the tightness of the theoretical results.**

***Index Terms*—distributed optimization, convex optimization, stochastic programming, stochastic gradient descent.**

I. Introduction

We consider the distributed optimization problem where a group of agents $\mathcal{N} = \{1, 2, \ldots, n\}$ collaboratively seek an $x \in \mathbb{R}^p$ that minimizes the average of n cost functions:

$$\min_{x\in\mathbb{R}^p} f(x) \left(= \frac{1}{n}\sum_{i=1}^{n} f_i(x)\right). \tag{1}$$

Each local cost function $f_i : \mathbb{R}^p \to \mathbb{R}$ is strongly convex, with Lipschitz continuous gradient, and is known by agent i only. The agents communicate and exchange information over a network. Problems in the form of (1) find applications in multi-agent target seeking [1], distributed machine learning [2], [3], [4], [5], [6], [7], [8], and wireless networks [9], [10], [5], among other scenarios.

In order to solve (1), we assume that at each iteration $k \geq 0$, the algorithm we study is able to obtain noisy gradient estimates $g_i(x_i(k), \xi_i(k))$, where $x_i(k)$ is the input for agent i, that satisfy the following condition.

This work was partially supported by the NSF under grants IIS-1914792, DMS-1664644, CNS-1645681, and ECCS-1933027, by the ONR under MURI grant N00014-19-1-2571, by the NIH under grants R01 GM135930 and UL54 TR004130, by the DOE under grants DE-AR-0001282 and DE-EE0009696, by the Boston University Kilachand Fund for Integrated Life Science and Engineering, by the Shenzhen Research Institute of Big Data (SRIBD) under grant J00120190011, and by the NSFC under grant 62003287.

S. Pu is with the School of Data Science, Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen, China. (e-mail: pushi@cuhk.edu.cn).

A. Olshevsky and I. Ch. Paschalidis are with the Department of Electrical and Computer Engineering and the Division of Systems Engineering, Boston University, Boston, MA (e-mails: alexols@bu.edu, yannisp@bu.edu).

Assumption 1. *For all $k \geq 0$, each random vector $\xi_i(k) \in \mathbb{R}^m$ is independent across $i \in \mathcal{N}$. Denote by $\mathcal{F}(k)$ the σ-algebra generated by $\{x_i(0), x_i(1), \ldots, x_i(k) \mid i \in \mathcal{N}\}$. Then,*

$$\begin{aligned} &\mathbb{E}_{\xi_i(k)}[g_i(x_i(k), \xi_i(k)) \mid \mathcal{F}(k)] = \nabla f_i(x_i(k)), \\ &\mathbb{E}_{\xi_i(k)}[\|g_i(x_i(k), \xi_i(k)) - \nabla f_i(x_i(k))\|^2 \mid \mathcal{F}(k)] \\ &\qquad \leq \sigma^2 + M\|\nabla f_i(x_i(k))\|^2, \quad \textit{for some } \sigma, M > 0. \end{aligned} \tag{2}$$

Stochastic gradients appear in many machine learning problems. For example, suppose $f_i(x) := \mathbb{E}_{\xi_i\sim\mathcal{D}_i}[F_i(x, \xi_i)]$ represents the expected loss function for agent i, where ξ_i are independent data samples gathered over time, and $\mathcal{D}_i$ represents the data distribution. Then for any x and ξ_i sampled from $\mathcal{D}_i$, $g_i(x, \xi_i) := \nabla F_i(x, \xi_i)$ is an unbiased estimator of $\nabla f_i(x)$. For another example, suppose $f_i(x) := (1/|\mathcal{S}_i|)\sum_{\zeta_j\in\mathcal{S}_i} F(x, \zeta_j)$ denotes an empirical risk function, where $\mathcal{S}_i$ is the local dataset for agent i. In this setting, the gradient estimation of $f_i(x)$ can incur noise from various sources such as minibatch random sampling of the local dataset and discretization for reducing communication cost [11].

Problem (1) has been studied extensively in the literature under various distributed algorithms [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], among which the distributed gradient descent (DGD) method proposed in [13] has drawn the greatest attention. Recently, distributed implementation of stochastic gradient algorithms has received considerable interest. Several works have shown that distributed methods may compare with their centralized counterparts under certain conditions. For example, the work in [24], [25], [26] first showed that, with sufficiently small constant stepsize, a distributed stochastic gradient method achieves comparable performance to a centralized method in terms of the steady-state mean-square-error.

Despite the aforementioned efforts, it is unclear how long, or how many iterations it takes for a distributed stochastic gradient method to reach the convergence rate of centralized SGD. The number of required iterations, called "transient time" of the algorithm, is a key measurement of the performance of the distributed implementation. In this work, we perform a non-asymptotic analysis for the distributed stochastic gradient descent (DSGD) method adapted from DGD and the diffusion strategy [1].[1] In addition to showing that in expectation, the algorithm asymptotically achieves the optimal convergence rate enjoyed by a centralized scheme, we precisely identify its non-asymptotic convergence rate as a function of characteristics

[1]Note that in [1] this method was called "Adapt-then-Combine".

of the objective functions and the network (e.g., spectral gap $(1-\rho_w)$ of the mixing matrix). Furthermore, we characterize the transient time needed for DSGD to achieve the optimal rate of convergence, which behaves as $\mathcal{O}\left(\frac{n}{(1-\rho_w)^2}\right)$ assuming certain conditions on the objective functions, stepsize policy and initial solutions. Finally, we construct a "hard" optimization problem for which we show the transient time needed for DSGD to approach the asymptotic convergence rate is lower bounded by $\Omega\left(\frac{n}{(1-\rho_w)^2}\right)$, implying the obtained transient time is sharp. These results are new to the best of our knowledge.

A. Related Works

We briefly discuss the related literature on (distributed) stochastic optimization. First of all, our work is related to stochastic approximation (SA) methods dating back to the seminal works [27] and [28]. For a strongly convex objective function f with Lipschitz continuous gradients, it has been shown that the optimal convergence rate for solving problem (1) is $\mathcal{O}(\frac{1}{k})$ under a diminishing stepsize policy [29].

Distributed stochastic gradient methods have received much attention in the recent years. For nonsmooth convex objective functions, the work in [30] considered distributed constrained optimization and established asymptotic convergence to the optimal set using two diminishing stepsizes to account for communication noise and subgradient errors, respectively. The paper [31] proposed a distributed dual averaging method which exhibits a convergence rate of $\mathcal{O}(\frac{n\log k}{(1-\lambda_2(\mathbf{W}))\sqrt{k}})$ under a carefully chosen SA stepsize sequence, where $\lambda_2(\mathbf{W})$ is the second largest singular value of the mixing matrix $\mathbf{W}$. A projected stochastic gradient algorithm was considered in [32] for solving nonconvex optimization problems by combining a local stochastic gradient update and a gossip step. This work proved that consensus is asymptotically achieved and the solutions converge to the set of KKT points with SA stepsizes. In [33], the authors proposed an adaptive diffusion algorithm based on penalty methods and showed that the expected optimization error is bounded by $\mathcal{O}(\alpha)$ under a constant stepsize α. The work in [34] considered distributed constrained convex optimization under multiple noise terms in both computation and communication stages. By means of an augmented Lagrangian framework, almost sure convergence with a diminishing stepsize policy was established. [35] investigated a subgradient-push method for distributed optimization over time-varying directed graphs. For strongly convex objective functions, the method exhibits an $\mathcal{O}(\frac{\ln k}{k})$ convergence rate. The work in [36] used a time-dependent weighted mixing of stochastic subgradient updates to achieve an $\mathcal{O}(\frac{n\sqrt{n}}{(1-\lambda_2(\mathbf{W}))k})$ convergence rate for minimizing the sum of nonsmooth strongly convex functions. [37] presented a new class of distributed first-order methods for nonsmooth and stochastic optimization which was shown to exhibit an $\mathcal{O}(\frac{1}{k})$ (respectively, $\mathcal{O}(\frac{1}{\sqrt{k}})$) convergence rate for minimizing the sum of strongly convex functions (respectively, convex functions). The work in [38] considered a decentralized algorithm with delayed gradient information which achieves an $\mathcal{O}(\frac{1}{\sqrt{k}})$ rate of convergence for general convex functions. In [39], an $\mathcal{O}(\frac{1}{k})$ convergence rate was established for strongly convex costs and random networks. Recently, the work in [40] proposed a variance-reduced decentralized stochastic optimization method with gradient tracking.

Several recent works have shown that distributed methods may compare with centralized algorithms under various conditions. In addition to [24], [25], [26] discussed before, [41], [42] proved that distributed stochastic approximation performs asymptotically as well as centralized schemes by means of a central limit theorem. [43] first showed that a distributed stochastic gradient algorithm asymptotically achieves comparable convergence rate to a centralized method, but assuming that all the local functions f_i have the same minimum. [44], [45] demonstrated the advantage of distributively implementing a stochastic gradient method assuming that sampling times are random and non-negligible. For nonconvex objective functions, [46] proved that decentralized algorithms can achieve a linear speedup similar to a centralized algorithm when k is large enough. This result was generalized to the setting of directed communication networks in [47] for training deep neural networks. The work in [48] considered a distributed stochastic gradient tracking method which performs as well as centralized stochastic gradient descent under a small enough constant stepsize. A recent paper [49] discussed an algorithm that asymptotically performs as well as the best bounds on centralized stochastic gradient descent subject to possible message losses, delays, and asynchrony. In a parallel recent work [50], a similar result was demonstrated with a further compression technique which allowed nodes to save on communication. For more discussion on the topic of achieving asymptotic network independence in distributed stochastic optimization, the readers are referred to a recent survey [51] .

B. Main Contribution

We next summarize the main contribution of the paper. First, we begin by performing a non-asymptotic convergence analysis of the distributed stochastic gradient descent (DSGD) method. For strongly convex and smooth objective functions, in expectation, DSGD asymptotically achieves the optimal network independent convergence rate compared to centralized stochastic gradient descent (SGD). We explicitly identify the non-asymptotic convergence rate as a function of characteristics of the objective functions and the network. The relevant results are established in Corollary 1 and Theorem 1.

Our main contribution is to characterize the transient time needed for DSGD to approach the asymptotic convergence rate. On the one hand, we show an upper bound of $K_T = \mathcal{O}\left(\frac{n}{(1-\rho_w)^2}\right)$, where $1-\rho_w$ denotes the spectral gap of the mixing matrix of communicating agents. On the other hand, we construct a "hard" optimization problem for which we show that the transient time needed for DSGD to approach the asymptotic convergence rate is lower bounded by $\Omega\left(\frac{n}{(1-\rho_w)^2}\right)$, implying that this upper bound is sharp.

Additionally, we provide numerical experiments that demonstrate the tightness of the theoretical findings. In particular, for the ring network topology and the square grid network topology, simulations are consistent with the transient time

$K_T = \mathcal{O}\left(\frac{n}{(1-\rho_w)^2}\right)$ for solving the *on-line* ridge regression problem.

C. Notation

Vectors are column vectors unless otherwise specified. Each agent i holds a local copy of the decision vector denoted by $x_i \in \mathbb{R}^p$, and its value at iteration/time k is written as $x_i(k)$. Let

$$\mathbf{x} := [x_1, x_2, \ldots, x_n]^\intercal \in \mathbb{R}^{n\times p}, \quad \overline{x} := \frac{1}{n}\sum_{i=1}^n x_i,$$

where $\intercal$ denotes transpose. Define an aggregate objective function

$$F(\mathbf{x}) := \sum_{i=1}^n f_i(x_i),$$

and let

$$\nabla F(\mathbf{x}) := [\nabla f_1(x_1), \nabla f_2(x_2), \ldots, \nabla f_n(x_n)]^\intercal \in \mathbb{R}^{n\times p},$$
$$\bar{\nabla} F(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^n \nabla f_i(x_i).$$

In addition, we denote

$$\boldsymbol{\xi} := [\xi_1, \xi_2, \ldots, \xi_n]^\intercal \in \mathbb{R}^{n\times p},$$
$$\mathbf{g}(\mathbf{x}, \boldsymbol{\xi}) := [g_1(x_1, \xi_1), g_2(x_2, \xi_2), \ldots, g_n(x_n, \xi_n)]^\intercal \in \mathbb{R}^{n\times p}.$$

In what follows we write $g_i(k) := g_i(x_i(k), \xi_i(k))$ and $\mathbf{g}(k) := \mathbf{g}(\mathbf{x}(k), \boldsymbol{\xi}(k))$ for short.

The inner product of two vectors a, b is written as $\langle a, b\rangle$. For two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times p}$, let $\langle \mathbf{A}, \mathbf{B}\rangle := \sum_{i=1}^n \langle A_i, B_i\rangle$, where A_i (respectively, B_i) is the i-th row of $\mathbf{A}$ (respectively, $\mathbf{B}$). We use $\|\cdot\|$ to denote the 2-norm of vectors and the Frobenius norm of matrices.

A graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ has a set of vertices (nodes) $\mathcal{N} = \{1, 2, \ldots, n\}$ and a set of edges connecting vertices $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$. Suppose agents interact in an undirected graph, i.e., $(i, j) \in \mathcal{E}$ if and only if $(j, i) \in \mathcal{E}$. Each agent i has a set of neighbors $\mathcal{N}_i = \{j \mid j \neq i, (i, j) \in \mathcal{E}\}$.

Denote the mixing matrix of agents by $\mathbf{W} = [w_{ij}] \in \mathbb{R}^{n\times n}$. Two agents i and j are connected if and only if $w_{ij}, w_{ji} > 0$ ($w_{ij} = w_{ji} = 0$ otherwise). Formally, we make the following assumption on the communication among agents.

Assumption 2. *The graph $\mathcal{G}$ is undirected and connected (there exists a path between any two nodes). The mixing matrix $\mathbf{W}$ is nonnegative and doubly stochastic, i.e., $\mathbf{W}\mathbf{1} = \mathbf{1}$ and $\mathbf{1}^\intercal\mathbf{W} = \mathbf{1}^\intercal$, where $\mathbf{1}$ denotes the vector of all ones.*

From Assumption 2, we have the following contraction property of $\mathbf{W}$ (see [20]).

Lemma 1. *Let Assumption 2 hold, and let ρ_w denote the spectral norm of the matrix $\mathbf{W} - \frac{1}{n}\mathbf{1}\mathbf{1}^\intercal$. Then, $\rho_w < 1$ and*

$$\|\mathbf{W}\boldsymbol{\omega} - \mathbf{1}\overline{\omega}\| \le \rho_w \|\boldsymbol{\omega} - \mathbf{1}\overline{\omega}\|$$

for all $\boldsymbol{\omega} \in \mathbb{R}^{n\times p}$, where $\overline{\omega} := \frac{1}{n}\mathbf{1}^\intercal\boldsymbol{\omega}$.

The rest of this paper is organized as follows. We present the DSGD algorithm and some preliminary results in Section II. In Section III we prove the sublinear convergence rate of the algorithm. Our main convergence results and a comparison with the centralized stochastic gradient method are in Section IV. Two numerical example are presented in Section V, and we conclude the paper in Section VI.

II. Distributed Stochastic Gradient Descent

We consider the following DSGD method adapted from DGD and the diffusion strategy [1]: at each step $k \ge 0$, every agent i independently performs the update:

$$x_i(k+1) = \sum\nolimits_{j\in\mathcal{N}_i\cup\{i\}} w_{ij}\left(x_j(k) - \alpha_k g_j(k)\right), \quad (3)$$

where $\{\alpha_k\}$ is a sequence of non-increasing stepsizes. The particular choice of the stepsize sequence will be introduced in Section III. The initial vectors $x_i(0)$ are arbitrary for all $i \in \mathcal{N}$. Since $w_{ij} = 0$ if agent i and agent j are not connected in the network, we can rewrite (3) in the following compact form:

$$\mathbf{x}(k+1) = \mathbf{W}\left(\mathbf{x}(k) - \alpha_k \mathbf{g}(k)\right). \quad (4)$$

Throughout the paper, we make the following standing assumption regarding the objective functions f_i.[2] These assumptions are satisfied for many machine learning problems, such as linear regression, smooth support vector machine (SVM), logistic regression, and softmax regression.

Assumption 3. *Each $f_i : \mathbb{R}^p \to \mathbb{R}$ is μ-strongly convex with L-Lipschitz continuous gradients, i.e., for any $x, x' \in \mathbb{R}^p$,*

$$\begin{aligned}&\langle \nabla f_i(x) - \nabla f_i(x'), x - x'\rangle \ge \mu\|x - x'\|^2,\\ &\|\nabla f_i(x) - \nabla f_i(x')\| \le L\|x - x'\|.\end{aligned} \quad (5)$$

Under Assumption 3, Problem (1) has a unique optimal solution $x_* \in \mathbb{R}^p$, and the following result holds (see [20] Lemma 10).

Lemma 2. *For any $x \in \mathbb{R}^p$ and $\alpha \in (0, 2/L)$, we have*

$$\|x - \alpha\nabla f(x) - x_*\| \le \lambda\|x - x_*\|,$$

where $\lambda = \max(|1 - \alpha\mu|, |1 - \alpha L|)$.

Denote $\overline{g}(k) := \frac{1}{n}\sum_{i=1}^n g_i(k)$. The following two results are useful for our analysis.

Lemma 3. *Under Assumptions 1 and 3, for all $k \ge 0$,*

$$\mathbb{E}\left[\left\|\overline{g}(k) - \bar{\nabla}F(\mathbf{x}(k))\right\|^2 \mid \mathcal{F}(k)\right] \le \frac{2ML^2}{n^2}\|\mathbf{x}(k) - \mathbf{1}x_*^\intercal\|^2 + \frac{\bar{M}}{n},$$

where

$$\bar{M} := \frac{2M\sum_{i=1}^n \|\nabla f_i(x_*)\|^2}{n} + \sigma^2.$$

[2] The assumption can be generalized to the case where the agents have different μ and L.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3126253, IEEE Transactions on Automatic Control

Proof. By definitions of $\overline{g}(k)$, $\bar{\nabla}F(\mathbf{x}(k))$ and Assumption 1, we have

$$\begin{aligned}&\mathbb{E}\left[\left\|\overline{g}(k)-\bar{\nabla}F(\mathbf{x}(k))\right\|^2\right]\\&=\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n g_i(k)-\frac{1}{n}\sum_{i=1}^n\nabla f_i(x_i(k))\right\|^2\right]\\&=\frac{1}{n^2}\sum_{i=1}^n\mathbb{E}\left[\|g_i(k)-\nabla f_i(x_i(k))\|^2\right]\\&\leq\frac{\sigma^2}{n}+\frac{M\sum_{i=1}^n\|\nabla f_i(x_i(k))\|^2}{n^2}.\end{aligned}$$

Notice that $\|\nabla f_i(x_i(k))\|^2=\|\nabla f_i(x_i(k))-\nabla f_i(x_*)+\nabla f_i(x_*)\|^2\leq 2\|\nabla f_i(x_i(k))-\nabla f_i(x_*)\|^2+2\|\nabla f_i(x_*)\|^2\leq 2L^2\|x_i(k)-x_*\|^2+2\|\nabla f_i(x_*)\|^2$ from Assumption 3. We have

$$\begin{aligned}\mathbb{E}\left[\left\|\overline{g}(k)-\bar{\nabla}F(\mathbf{x}(k))\right\|^2\right]&\leq\frac{2ML^2}{n^2}\|\mathbf{x}(k)-\mathbf{1}x_*^{\intercal}\|^2\\&\quad+\frac{1}{n}\left(\frac{2M\sum_{i=1}^n\|\nabla f_i(x_*)\|^2}{n}+\sigma^2\right).\end{aligned}$$

□

Lemma 4. *Under Assumption 3, for all* $k\geq 0$*,*

$$\left\|\nabla f(\overline{x}(k))-\bar{\nabla}F(\mathbf{x}(k))\right\|\leq\frac{L}{\sqrt{n}}\|\mathbf{x}(k)-\mathbf{1}\overline{x}(k)^{\intercal}\|. \quad (6)$$

Proof. By definition,

$$\begin{aligned}&\left\|\nabla f(\overline{x}(k))-\bar{\nabla}F(\mathbf{x}(k))\right\|\\&=\left\|\frac{1}{n}\sum_{i=1}^n\nabla f_i(\overline{x}(k))-\frac{1}{n}\sum_{i=1}^n\nabla f_i(x_i(k))\right\|\\ \text{(Assumption 3)}\quad&\leq\frac{L}{n}\sum_{i=1}^n\|\overline{x}(k)-x_i(k)\|\\&\leq\frac{L}{\sqrt{n}}\|\mathbf{x}(k)-\mathbf{1}\overline{x}(k)^{\intercal}\|,\end{aligned}$$

where the last relation follows from Hölder's inequality. □

A. Preliminary Results

In this section, we present some preliminary results concerning $\mathbb{E}[\|\overline{x}(k)-x_*\|^2]$ (expected optimization error) and $\mathbb{E}[\|\mathbf{x}(k)-\mathbf{1}\overline{x}(k)^{\intercal}\|^2]$ (expected consensus error). Specifically, we bound the two terms by linear combinations of their values in the last iteration.

For ease of presentation, for all k we denote

$$U(k):=\mathbb{E}\left[\|\overline{x}(k)-x_*\|^2\right],\ V(k):=\mathbb{E}[\|\mathbf{x}(k)-\mathbf{1}\overline{x}(k)^{\intercal}\|^2].$$

In the lemma below, we bound the optimization error $U(k+1)$ by several error terms at iteration k, including the consensus error $V(k)$. It serves as a starting point for the follow-up analysis.

Lemma 5. *Suppose Assumptions 1-3 hold. Under Algorithm* (4)*, supposing* $\alpha_k\leq\frac{1}{L}$*, we have*

$$\begin{aligned}U(k+1)&\leq(1-\alpha_k\mu)^2U(k)\\&\quad+\frac{2\alpha_k L}{\sqrt{n}}(1-\alpha_k\mu)\mathbb{E}[\|\overline{x}(k)-x_*\|\|\mathbf{x}(k)-\mathbf{1}\overline{x}(k)^{\intercal}\|]\\&\quad+\frac{\alpha_k^2L^2}{n}V(k)+\alpha_k^2\left(\frac{2ML^2}{n^2}\mathbb{E}[\|\mathbf{x}(k)-\mathbf{1}x_*^{\intercal}\|^2]+\frac{\bar{M}}{n}\right).\end{aligned} \quad (7)$$

Proof. By the definitions of $\bar{x}(k)$, $\overline{g}(k)$ and relation (4), we have $\overline{x}(k+1)=\overline{x}(k)-\alpha_k\overline{g}(k)$. Hence,

$$\begin{aligned}&\|\overline{x}(k+1)-x_*\|^2=\|\overline{x}(k)-\alpha_k\overline{g}(k)-x_*\|^2\\=&\left\|\overline{x}(k)-\alpha_k\bar{\nabla}F(\mathbf{x}(k))-x_*+\alpha_k\bar{\nabla}F(\mathbf{x}(k))-\alpha_k\overline{g}(k)\right\|^2\\=&\|\overline{x}(k)-\alpha_k\bar{\nabla}F(\mathbf{x}(k))-x_*\|^2+\alpha_k^2\|\bar{\nabla}F(\mathbf{x}(k))-\overline{g}(k)\|^2\\&+2\alpha_k\langle\overline{x}(k)-\alpha_k\bar{\nabla}F(\mathbf{x}(k))-x_*,\bar{\nabla}F(\mathbf{x}(k))-\overline{g}(k)\rangle.\end{aligned}$$

Noting that $\mathbb{E}[\overline{g}(k)\mid\mathbf{x}(k)]=\bar{\nabla}F(\mathbf{x}(k))$ and $\mathbb{E}[\|\overline{g}(k)-\bar{\nabla}F(\mathbf{x}(k))\|^2\mid\mathcal{F}(k)]\leq\frac{2ML^2}{n^2}\|\mathbf{x}(k)-\mathbf{1}x_*^{\intercal}\|^2+\frac{\bar{M}}{n}$ from Lemma 3, we obtain

$$\begin{aligned}\mathbb{E}[\|\overline{x}(k+1)-x_*\|^2\mid\mathcal{F}(k)]&\leq\|\overline{x}(k)-\alpha_k\bar{\nabla}F(\mathbf{x}(k))-x_*\|^2\\&\quad+\alpha_k^2\left(\frac{2ML^2}{n^2}\|\mathbf{x}(k)-\mathbf{1}x_*^{\intercal}\|^2+\frac{\bar{M}}{n}\right).\end{aligned} \quad (8)$$

We next bound the first term on the right-hand-side of (8).

$$\begin{aligned}&\|\overline{x}(k)-\alpha_k\bar{\nabla}F(\mathbf{x}(k))-x_*\|^2\\=&\|\overline{x}(k)-\alpha_k\nabla f(\overline{x}(k))-x_*+\alpha_k\nabla f(\overline{x}(k))-\alpha_k\bar{\nabla}F(\mathbf{x}(k))\|^2\\\leq&\|\overline{x}(k)-\alpha_k\nabla f(\overline{x}(k))-x_*\|^2\\&+2\alpha_k\|\overline{x}(k)-\alpha_k\nabla f(\overline{x}(k))-x_*\|\|\nabla f(\overline{x}(k))-\bar{\nabla}F(\mathbf{x}(k))\|\\&+\alpha_k^2\|\nabla f(\overline{x}(k))-\bar{\nabla}F(\mathbf{x}(k))\|^2,\end{aligned}$$

where we used the Cauchy-Schwarz inequality. By Lemma 3,

$$\|\nabla f(\overline{x}(k))-\bar{\nabla}F(\mathbf{x}(k))\|^2\leq\tfrac{L^2}{n}\|\mathbf{x}(k)-\mathbf{1}\overline{x}(k)^{\intercal}\|^2.$$

Since $\alpha_k\leq\frac{1}{L}$, in light of Lemma 2,

$$\|\overline{x}(k)-\alpha_k\nabla f(\overline{x}(k))-x_*\|^2\leq(1-\alpha_k\mu)^2\|\overline{x}(k)-x_*\|^2.$$

Then we have

$$\begin{aligned}\|\overline{x}(k)-\alpha_k\bar{\nabla}F(\mathbf{x}(k))-x_*\|^2&\leq(1-\alpha_k\mu)^2\|\overline{x}(k)-x_*\|^2\\&\quad+\frac{2\alpha_k L}{\sqrt{n}}(1-\alpha_k\mu)\|\overline{x}(k)-x_*\|\|\mathbf{x}(k)-\mathbf{1}\overline{x}(k)^{\intercal}\|\\&\quad+\frac{\alpha_k^2L^2}{n}\|\mathbf{x}(k)-\mathbf{1}\overline{x}(k)^{\intercal}\|^2.\end{aligned} \quad (9)$$

In light of relation (9), taking full expectation on both sides of relation (8) yields the result. □

The next result is a corollary of Lemma 5 with an additional condition on the stepsize α_k. We are able to remove the cross term in relation (7) and obtain a cleaner expression, which facilitates our later analysis.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3126253, IEEE Transactions on Automatic Control

Lemma 6. *Suppose Assumptions 1-3 hold. Under Algorithm* (4)*, supposing* $\alpha_k \leq \min\{\frac{1}{L}, \frac{1}{3\mu}\}$*, then*

$$U(k+1) \leq \left(1 - \frac{3}{2}\alpha_k\mu\right)U(k) + \frac{3\alpha_k L^2}{n\mu}V(k) + \alpha_k^2\left(\frac{2ML^2}{n^2}\mathbb{E}[\|\mathbf{x}(k) - \mathbf{1}x_*^\intercal\|^2] + \frac{\bar{M}}{n}\right). \quad (10)$$

Proof. From Lemma 5,

$$\begin{aligned} U(k+1) &\leq (1-\alpha_k\mu)^2 U(k) + (1-\alpha_k\mu)^2 cU(k) \\ &+ \frac{\alpha_k^2 L^2}{n}\frac{1}{c}V(k) + \frac{\alpha_k^2 L^2}{n}V(k) \\ &+ \alpha_k^2\left(\frac{2ML^2}{n^2}\mathbb{E}[\|\mathbf{x}(k) - \mathbf{1}x_*^\intercal\|^2] + \frac{\bar{M}}{n}\right) \\ &\leq (1+c)(1-\alpha_k\mu)^2 U(k) + \left(1+\frac{1}{c}\right)\frac{\alpha_k^2 L^2}{n}V(k) \\ &+ \alpha_k^2\left(\frac{2ML^2}{n^2}\mathbb{E}[\|\mathbf{x}(k) - \mathbf{1}x_*^\intercal\|^2] + \frac{\bar{M}}{n}\right), \end{aligned}$$

where $c > 0$ is arbitrary.

Take $c = \frac{3}{8}\alpha_k\mu$. Noting that $\alpha_k \leq \frac{1}{3\mu}$, we have $(1+c)(1-\alpha_k\mu)^2 \leq 1 - \frac{3}{2}\alpha_k\mu$, and $(1+\frac{1}{c})\alpha_k \leq \frac{3}{\mu}$. Thus,

$$U(k+1) \leq \left(1 - \frac{3}{2}\alpha_k\mu\right)U(k) + \frac{3\alpha_k L^2}{n\mu}V(k) + \alpha_k^2\left(\frac{2ML^2}{n^2}\mathbb{E}[\|\mathbf{x}(k) - \mathbf{1}x_*^\intercal\|^2] + \frac{\bar{M}}{n}\right).$$

□

Since the consensus error term $V(k)$ plays a key role in the statements of Lemma 5 and Lemma 6, we present the following lemma that bounds $V(k+1)$.

Lemma 7. *Suppose Assumptions 1-3 hold. Under Algorithm* (4)*, for all* $k \geq 0$,

$$\begin{aligned} V(k+1) &\leq \frac{(3+\rho_w^2)}{4}V(k) + \alpha_k^2\rho_w^2 n\sigma^2 \\ &+ 2\alpha_k^2\rho_w^2\left(\frac{3}{1-\rho_w^2} + M\right)\left(L^2\mathbb{E}[\|\mathbf{x}(k) - \mathbf{1}x_*^\intercal\|^2]\right. \\ &\left.+\|\nabla F(\mathbf{1}x_*^\intercal)\|^2\right). \end{aligned} \quad (11)$$

Proof. From relation (4),

$$\begin{aligned} &\mathbf{x}(k+1) - \mathbf{1}\overline{x}(k+1)^\intercal \\ &= \mathbf{W}(\mathbf{x}(k) - \alpha_k\mathbf{g}(k)) - \mathbf{1}(\overline{x}(k) - \alpha_k\overline{g}(k)) \\ &= \left(\mathbf{W} - \frac{\mathbf{1}\mathbf{1}^\intercal}{n}\right)[(\mathbf{x}(k) - \mathbf{1}\overline{x}(k)^\intercal) - \alpha_k(\mathbf{g}(k) - \mathbf{1}\overline{g}(k)^\intercal)], \end{aligned}$$

we have

$$\begin{aligned} &\|\mathbf{x}(k+1) - \mathbf{1}\overline{x}(k+1)^\intercal\|^2 \\ &\leq \rho_w^2\,\|(\mathbf{x}(k) - \mathbf{1}\overline{x}(k)^\intercal) - \alpha_k(\mathbf{g}(k) - \mathbf{1}\overline{g}(k)^\intercal)\|^2 \\ &= \rho_w^2\Big[\|\mathbf{x}(k) - \mathbf{1}\overline{x}(k)^\intercal\|^2 + \alpha_k^2\|\mathbf{g}(k) - \mathbf{1}\overline{g}(k)^\intercal\|^2 \\ &\quad - 2\alpha_k\langle\mathbf{x}(k) - \mathbf{1}\overline{x}(k)^\intercal, \mathbf{g}(k) - \mathbf{1}\overline{g}(k)^\intercal\rangle\Big]. \end{aligned}$$

Since $\mathbb{E}[\mathbf{g}(k) \mid \mathcal{F}(k)] = \nabla F(\mathbf{x}(k))$ and $\mathbb{E}[\overline{g}(k) \mid \mathcal{F}(k)] = \bar{\nabla} F(\mathbf{x}(k))$,

$$\begin{aligned} &\mathbb{E}\left[\langle\mathbf{x}(k) - \mathbf{1}\overline{x}(k)^\intercal, \mathbf{g}(k) - \mathbf{1}\overline{g}(k)^\intercal\rangle \mid \mathcal{F}(k)\right] \\ &\quad = \langle\mathbf{x}(k) - \mathbf{1}\overline{x}(k)^\intercal, \nabla F(\mathbf{x}(k)) - \mathbf{1}\bar{\nabla}F(\mathbf{x}(k))\rangle, \end{aligned}$$

and

$$\begin{aligned} &\mathbb{E}[\|\mathbf{g}(k) - \mathbf{1}\overline{g}(k)^\intercal\|^2 \mid \mathcal{F}(k)] \\ &= \mathbb{E}[\|\nabla F(\mathbf{x}(k)) - \mathbf{1}\bar{\nabla}F(\mathbf{x}(k)) - \nabla F(\mathbf{x}(k)) + \mathbf{1}\bar{\nabla}F(\mathbf{x}(k)) \\ &\quad + \mathbf{g}(k) - \mathbf{1}\overline{g}(k)^\intercal\|^2 \mid \mathcal{F}(k)] \\ &= \|\nabla F(\mathbf{x}(k)) - \mathbf{1}\bar{\nabla}F(\mathbf{x}(k))\|^2 \\ &\quad + \mathbb{E}[\|\nabla F(\mathbf{x}(k)) - \mathbf{g}(k) - \\ &\qquad (\mathbf{1}\bar{\nabla}F(\mathbf{x}(k)) - \mathbf{1}\overline{g}(k)^\intercal)\|^2 \mid \mathcal{F}(k)] \\ &\leq \|\nabla F(\mathbf{x}(k)) - \mathbf{1}\bar{\nabla}F(\mathbf{x}(k))\|^2 \\ &\quad + \mathbb{E}[\|\nabla F(\mathbf{x}(k)) - \mathbf{g}(k)\|^2 \mid \mathcal{F}(k)] \\ &\leq \|\nabla F(\mathbf{x}(k)) - \mathbf{1}\bar{\nabla}F(\mathbf{x}(k))\|^2 + n\sigma^2 + M\|\nabla F(\mathbf{x}(k))\|^2, \end{aligned}$$

where the last inequality follows from Assumption 1. Therefore (assuming $\rho_w > 0$),

$$\begin{aligned} &\frac{1}{\rho_w^2}\mathbb{E}[\|\mathbf{x}(k+1) - \mathbf{1}\overline{x}(k+1)^\intercal\|^2 \mid \mathcal{F}(k)] \\ &\leq \|\mathbf{x}(k) - \mathbf{1}\overline{x}(k)^\intercal\|^2 + \alpha_k^2\|\nabla F(\mathbf{x}(k)) - \mathbf{1}\bar{\nabla}F(\mathbf{x}(k))\|^2 \\ &\quad + \alpha_k^2\left(n\sigma^2 + M\|\nabla F(\mathbf{x}(k))\|^2\right) \\ &\quad - 2\alpha_k\langle\mathbf{x}(k) - \mathbf{1}\overline{x}(k)^\intercal, \nabla F(\mathbf{x}(k)) - \mathbf{1}\bar{\nabla}F(\mathbf{x}(k))\rangle \\ &\leq \|\mathbf{x}(k) - \mathbf{1}\overline{x}(k)^\intercal\|^2 + \alpha_k^2(1+M)\|\nabla F(\mathbf{x}(k))\|^2 + \alpha_k^2 n\sigma^2 \\ &\quad + 2\alpha_k\,\|\mathbf{x}(k) - \mathbf{1}\overline{x}(k)^\intercal\|\,\|\nabla F(\mathbf{x}(k))\| \\ &\leq (1+c)\,\|\mathbf{x}(k) - \mathbf{1}\overline{x}(k)^\intercal\|^2 + \alpha_k^2(1+M)\|\nabla F(\mathbf{x}(k))\|^2 \\ &\quad + \alpha_k^2 n\sigma^2 + \frac{\alpha_k^2}{c}\,\|\nabla F(\mathbf{x}(k))\|^2, \end{aligned}$$

where $c > 0$ is arbitrary. Letting $c = \frac{1-\rho_w^2}{2}$ and noting that by Assumption 3,

$$\begin{aligned} &\|\nabla F(\mathbf{x}(k))\|^2 \leq 2\|\nabla F(\mathbf{x}(k)) - \nabla F(\mathbf{1}x_*^\intercal)\|^2 + 2\|\nabla F(\mathbf{1}x_*^\intercal)\|^2 \\ &\leq 2L^2\|\mathbf{x}(k) - \mathbf{1}x_*^\intercal\|^2 + 2\|\nabla F(\mathbf{1}x_*^\intercal)\|^2, \end{aligned}$$

we have

$$\begin{aligned} &\frac{1}{\rho_w^2}\mathbb{E}[\|\mathbf{x}(k+1) - \mathbf{1}\overline{x}(k+1)^\intercal\|^2 \mid \mathcal{F}(k)] - \alpha_k^2 n\sigma^2 \\ &\leq \left(\frac{3-\rho_w^2}{2}\right)\|\mathbf{x}(k) - \mathbf{1}\overline{x}(k)^\intercal\|^2 \\ &\quad + 2\alpha_k^2\left(\frac{3}{1-\rho_w^2} + M\right)(L^2\|\mathbf{x}(k) - \mathbf{1}x_*^\intercal\|^2 + \|\nabla F(\mathbf{1}x_*^\intercal)\|^2). \end{aligned}$$

Notice that $\rho_w^2\left(\frac{3-\rho_w^2}{2}\right) \leq \frac{(3+\rho_w^2)}{4}$. In light of Lemma 8, taking full expectation on both sides of the above inequality leads to

$$\begin{aligned} &\mathbb{E}[\|\mathbf{x}(k+1) - \mathbf{1}\overline{x}(k+1)^\intercal\|^2] - \alpha_k^2\rho_w^2 n\sigma^2 \\ &\leq \frac{(3+\rho_w^2)}{4}\mathbb{E}[\|\mathbf{x}(k) - \mathbf{1}\overline{x}(k)^\intercal\|^2] \\ &\quad + 2\alpha_k^2\rho_w^2\left(\frac{3}{1-\rho_w^2} + M\right)\left(L^2\mathbb{E}[\|\mathbf{x}(k) - \mathbf{1}x_*^\intercal\|^2]\right. \\ &\quad \left.+\|\nabla F(\mathbf{1}x_*^\intercal)\|^2\right). \end{aligned}$$

□

III. Analysis

We are now ready to derive some preliminary convergence results for Algorithm (4). First, we provide a uniform bound on the iterates generated by Algorithm (4) (in expectation) for all $k \geq 0$. Then based on the lemmas established in Section II-A, we prove the sublinear convergence rate for Algorithm (4), i.e., $U(k) = \mathcal{O}(\frac{1}{k})$ and $V(k) = \mathcal{O}(\frac{1}{k^2})$. These results provide the foundation for our main convergence theorems in Section IV.

From now on we consider the following stepsize policy:

$$\alpha_k := \frac{\theta}{\mu(k+K)}, \quad \forall k, \tag{12}$$

where constant $\theta > 1$, and

$$K := \left\lceil \frac{2\theta(1+M)L^2}{\mu^2} \right\rceil, \tag{13}$$

with $\lceil \cdot \rceil$ denoting the ceiling function.

A. Uniform Bound

We first derive a uniform bound on the iterates generated by Algorithm (4) (in expectation) for all $k \geq 0$. Such a result is helpful for bounding the error terms on the right hand sides of (7), (10) and (11).

Lemma 8. *Suppose Assumptions 1-3 hold. Under Algorithm* (4) *with stepsize policy* (12)*, for all* $k \geq 0$*, we have*

$$\mathbb{E}[\|\mathbf{x}(k) - \mathbf{1}x_*^\intercal\|^2] \leq \hat{X} := \max\left\{\|\mathbf{x}(0) - \mathbf{1}x_*^\intercal\|^2,\ \frac{9\|\nabla F(\mathbf{1}x_*^\intercal)\|^2}{\mu^2} + \frac{n\sigma^2}{(1+M)L^2}\right\}. \tag{14}$$

Proof. The following arguments are inspired by those in [35].

First, we bound $\mathbb{E}\left[\|x_i(k) - \alpha_k g_i(k)\|^2\right]$ for all $i \in \mathcal{N}$ and $k \geq 0$. By Assumption 1,

$$\begin{aligned}
&\mathbb{E}\left[\|x_i(k) - \alpha_k g_i(k)\|^2 \mid \mathcal{F}(k)\right] \\
&= \|x_i(k) - \alpha_k \nabla f_i(x_i(k))\|^2 \\
&\quad + \alpha_k^2 \mathbb{E}\left[\|\nabla f_i(x_i(k)) - g_i(k)\|^2 \mid \mathcal{F}(k)\right] \\
&\leq \|x_i(k)\|^2 - 2\alpha_k \langle \nabla f_i(x_i(k)), x_i(k)\rangle + \alpha_k^2 \|\nabla f_i(x_i(k))\|^2 \\
&\quad + \alpha_k^2(\sigma^2 + M\|\nabla f_i(x_i(k))\|^2).
\end{aligned}$$

From the strong convexity and Lipschitz continuity of f_i, we know that

$$\begin{aligned}
&\langle \nabla f_i(x_i(k)), x_i(k)\rangle \\
&= \langle \nabla f_i(x_i(k)) - \nabla f_i(0), x_i(k) - 0\rangle + \langle \nabla f_i(0), x_i(k)\rangle \\
&\geq \mu\|x_i(k)\|^2 + \langle \nabla f_i(0), x_i(k)\rangle,
\end{aligned}$$

and

$$\begin{aligned}
\|\nabla f_i(x_i(k))\|^2 &= \|\nabla f_i(x_i(k)) - \nabla f_i(0) + \nabla f_i(0)\|^2 \\
&\leq 2L^2\|x_i(k)\|^2 + 2\|\nabla f_i(0)\|^2.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\mathbb{E}\left[\|x_i(k) - \alpha_k g_i(k)\|^2 \mid \mathcal{F}(k)\right] \\
&\leq \|x_i(k)\|^2 - 2\alpha_k\left[\mu\|x_i(k)\|^2 + \langle \nabla f_i(0), x_i(k)\rangle\right] \\
&\quad + 2\alpha_k^2(1+M)\left(L^2\|x_i(k)\|^2 + \|\nabla f_i(0)\|^2\right) + \alpha_k^2\sigma^2 \\
&\leq \|x_i(k)\|^2 - 2\alpha_k\mu\|x_i(k)\|^2 + 2\alpha_k\|\nabla f_i(0)\|\|x_i(k)\| \\
&\quad + 2\alpha_k^2(1+M)\left(L^2\|x_i(k)\|^2 + \|\nabla f_i(0)\|^2\right) + \alpha_k^2\sigma^2 \\
&\leq [1 - 2\alpha_k\mu + 2\alpha_k^2(1+M)L^2]\|x_i(k)\|^2 \\
&\quad + 2\alpha_k\|\nabla f_i(0)\|\|x_i(k)\| + \alpha_k^2\left[2(1+M)\|\nabla f_i(0)\|^2 + \sigma^2\right].
\end{aligned}$$

Taking full expectation on both sides, it follows that

$$\begin{aligned}
&\mathbb{E}\left[\|x_i(k) - \alpha_k g_i(k)\|^2\right] \\
&\leq [1 - 2\alpha_k\mu + 2\alpha_k^2(1+M)L^2]\mathbb{E}[\|x_i(k)\|^2] \\
&\quad + 2\alpha_k\|\nabla f_i(0)\|\sqrt{\mathbb{E}[\|x_i(k)\|^2]} \\
&\quad + \alpha_k^2\left[2(1+M)\|\nabla f_i(0)\|^2 + \sigma^2\right].
\end{aligned}$$

From the definition of K in (13), $\alpha_k \leq \frac{\mu}{2(1+M)L^2}$ for all $k \geq 0$. Hence,

$$\begin{aligned}
&\mathbb{E}\left[\|x_i(k) - \alpha_k g_i(k)\|^2\right] \\
&\leq (1 - \alpha_k\mu)\mathbb{E}[\|x_i(k)\|^2] + 2\alpha_k\|\nabla f_i(0)\|\sqrt{\mathbb{E}[\|x_i(k)\|^2]} \\
&\quad + \alpha_k^2\left[2(1+M)\|\nabla f_i(0)\|^2 + \sigma^2\right] \\
&\leq \mathbb{E}[\|x_i(k)\|^2] - \alpha_k\Big[\mu\mathbb{E}[\|x_i(k)\|^2] - 2\|\nabla f_i(0)\|\sqrt{\mathbb{E}[\|x_i(k)\|^2]} \\
&\quad - \frac{\mu}{2L^2}\left(2\|\nabla f_i(0)\|^2 + \frac{\sigma^2}{(1+M)}\right)\Big].
\end{aligned} \tag{15}$$

Let us define the following set:

$$\mathcal{X}_i := \left\{q \geq 0 : \mu q - 2\|\nabla f_i(0)\|\sqrt{q} - \frac{\mu}{2L^2}\left(2\|\nabla f_i(0)\|^2 + \frac{\sigma^2}{(1+M)}\right) \leq 0\right\}, \tag{16}$$

which is non-empty and compact. If $\mathbb{E}[\|x_i(k)\|^2] \notin \mathcal{X}_i$, we know from inequality (15) that $\mathbb{E}\left[\|x_i(k) - \alpha_k g_i(k)\|^2\right] \leq \mathbb{E}[\|x_i(k)\|^2]$. Otherwise,

$$\begin{aligned}
&\mathbb{E}\left[\|x_i(k) - \alpha_k g_i(k)\|^2\right] \leq \max_{q\in\mathcal{X}_i}\Big\{q - \frac{\mu}{2(1+M)L^2}\Big[\mu q \\
&\quad -2\|\nabla f_i(0)\|\sqrt{q} - \frac{\mu}{2L^2}\left(2\|\nabla f_i(0)\|^2 + \frac{\sigma^2}{(1+M)}\right)\Big]\Big\} \\
&= \max_{q\in\mathcal{X}_i}\Big\{\left(1 - \frac{\mu^2}{2(1+M)L^2}\right)q + \frac{\mu}{(1+M)L^2}\|\nabla f_i(0)\|\sqrt{q} \\
&\quad + \frac{\mu^2}{4(1+M)L^4}\left(2\|\nabla f_i(0)\|^2 + \frac{\sigma^2}{(1+M)}\right)\Big\}.
\end{aligned}$$

Define the last term above as R_i. The previous arguments imply that for all $k \geq 0$,

$$\mathbb{E}\left[\|x_i(k) - \alpha_k g_i(k)\|^2\right] \leq \max\{\mathbb{E}\left[\|x_i(k)\|^2\right], R_i\}.$$

Note that from relation (4),

$$\begin{aligned}
\|\mathbf{x}(k+1)\|^2 &\leq \|\mathbf{W}\|^2\|\mathbf{x}(k) - \alpha_k\mathbf{g}(k)\|^2 \\
&\leq \|\mathbf{x}(k) - \alpha_k\mathbf{g}(k)\|^2.
\end{aligned}$$

We have

$$\mathbb{E}[\|\mathbf{x}(k)\|^2] \leq \max\left\{\|\mathbf{x}(0)\|^2, \textstyle\sum_{i=1}^n R_i\right\}. \tag{17}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3126253, IEEE Transactions on Automatic Control

We further bound R_i as follows. From the definition of $\mathcal{X}_i$,

$$\max_{q\in\mathcal{X}_i} q \leq \frac{8\|\nabla f_i(0)\|^2}{\mu^2} + \frac{3\sigma^2}{4(1+M)L^2}.$$

Hence,

$$\begin{aligned}
R_i &= \max_{q\in\mathcal{X}_i}\Big\{q - \tfrac{\mu}{2(1+M)L^2}\Big[\mu q - 2\|\nabla f_i(0)\|\sqrt{q} \\
&\quad - \tfrac{\mu}{2L^2}\Big(2\|\nabla f_i(0)\|^2 + \tfrac{\sigma^2}{(1+M)}\Big)\Big]\Big\} \\
&\leq \max_{q\in\mathcal{X}_i} q - \tfrac{\mu}{2(1+M)L^2}\min_{q\in\mathcal{X}_i}\Big\{\mu q - 2\|\nabla f_i(0)\|\sqrt{q} \\
&\quad - \tfrac{\mu}{2L^2}\Big(2\|\nabla f_i(0)\|^2 + \tfrac{\sigma^2}{(1+M)}\Big)\Big\} \\
&\leq \tfrac{8\|\nabla f_i(0)\|^2}{\mu^2} + \tfrac{3\sigma^2}{4(1+M)L^2} \\
&\quad + \tfrac{\mu}{2(1+M)L^2}\Big[\tfrac{\|\nabla f_i(0)\|^2}{\mu} + \tfrac{\mu}{2L^2}\Big(2\|\nabla f_i(0)\|^2 + \tfrac{\sigma^2}{(1+M)}\Big)\Big] \\
&\leq \tfrac{9\|\nabla f_i(0)\|^2}{\mu^2} + \tfrac{\sigma^2}{(1+M)L^2}.
\end{aligned} \tag{18}$$

In light of inequality (18), further noticing that the choice of 0 is arbitrary in the proof of (17), we obtain the uniform bound for $\mathbb{E}[\|\mathbf{x}(k) - \mathbf{1}x_*^\intercal\|^2]$ in (14). □

The uniform bound provided in Lemma 8 is critical for deriving the sublinear convergence rates of $U(k)$ and $V(k)$, as it holds for all $k \geq 0$.

B. Sublinear Rate

With the help of Lemma 6 and Lemma 7 from Section II-A and Lemma 8, we show in Lemma 10 below that Algorithm (4) enjoys the sublinear convergence rate, i.e., $U(k) = \mathcal{O}(\frac{1}{k})$ and $V(k) = \mathcal{O}(\frac{1}{k^2})$. For the ease of analysis, we define two auxiliary variables:

$$\tilde{U}(k) := U(k-K),\ \tilde{V}(k) := V(k-K),\ \forall k \geq K. \tag{19}$$

We first derive uniform upper bounds for $U(k)$ and $V(k)$ respectively for all $k \geq 0$ based on Lemma 8. With these bounds, we are able to characterize the constants appearing in the sublinear convergence rates for $U(k)$ and $V(k)$ in Lemma 10 and Lemma 12 respectively.

Lemma 9. *Suppose Assumptions 1-3 hold. Under Algorithm* (4)*, we have*

$$U(k) \leq \frac{\hat{X}}{n},\ V(k) \leq \hat{X},\ \forall k \geq 0. \tag{20}$$

Proof. By definitions of $U(k)$, $V(k)$, and Lemma 8, we have

$$\begin{aligned}
U(k) &= \mathbb{E}[\|\overline{x}(k) - x_*\|^2] \\
&\leq \frac{1}{n}\mathbb{E}[\|\mathbf{x}(k) - \mathbf{1}x_*^\intercal\|^2] \leq \frac{\hat{X}}{n}, \\
V(k) &= \mathbb{E}[\|\mathbf{x}(k) - \mathbf{1}\overline{x}(k)^\intercal\|^2] \\
&\leq \mathbb{E}[\|\mathbf{x}(k) - \mathbf{1}x_*^\intercal\|^2] \leq \hat{X}.
\end{aligned}$$

□

Denote an auxiliary counter

$$\tilde{k} := k + K, \quad \forall k \geq 0. \tag{21}$$

Our strategy is to first show that the consensus error of Algorithm (4) decays as $V(k) = \mathcal{O}(\frac{1}{\tilde{k}^2})$ based on Lemma 7, since $U(k)$ does not appear explicitly in relation (11).

Lemma 10. *Suppose Assumptions 1-3 hold. Let*

$$K_1 := \left\lceil \max\left\{2K, \frac{16}{1-\rho_w^2}\right\}\right\rceil. \tag{22}$$

Under Algorithm (4) *with stepsize* (12)*, for all $k \geq K_1 - K$, we have*

$$V(k) \leq \frac{\hat{V}}{\tilde{k}^2},$$

where

$$\hat{V} := \max\left\{K_1^2\hat{X}, \frac{8\theta^2\rho_w^2 c_1}{\mu^2(1-\rho_w^2)}\right\}, \tag{23}$$

with

$$c_1 := 2\left(\frac{3}{1-\rho_w^2} + M\right)\left(L^2\hat{X} + \|\nabla F(\mathbf{1}x_*^\intercal)\|^2\right) + n\sigma^2. \tag{24}$$

Proof. From Lemma 7 and Lemma 8, for $k \geq 0$,

$$V(k+1) \leq \frac{(3+\rho_w^2)}{4}V(k) + \alpha_k^2\rho_w^2 c_1, \tag{25}$$

with c_1 defined in (24). From the definitions of α_k and $\tilde{V}(k)$ in (12) and (19) respectively, we know that when $k \geq K$,

$$\tilde{V}(k+1) \leq \frac{(3+\rho_w^2)}{4}\tilde{V}(k) + \frac{\theta^2\rho_w^2 c_1}{\mu^2}\left(\frac{1}{k^2}\right).$$

We now prove the lemma by induction. For $k = K_1$, we know $\tilde{V}(k) = \frac{K_1^2\tilde{V}(K_1)}{K_1^2} \leq \frac{\hat{V}}{k^2}$ from Lemma 9. Now suppose $\tilde{V}(k) \leq \frac{\hat{V}}{k^2}$ for some $k \geq K_1$, then

$$\tilde{V}(k+1) \leq \frac{(3+\rho_w^2)}{4}\frac{\hat{V}}{k^2} + \frac{\theta^2\rho_w^2 c_1}{\mu^2}\left(\frac{1}{k^2}\right).$$

To show that $\tilde{V}(k+1) \leq \frac{\hat{V}}{(k+1)^2}$, it is sufficient to show

$$\frac{(3+\rho_w^2)}{4}\frac{\hat{V}}{k^2} + \frac{\theta^2\rho_w^2 c_1}{\mu^2}\left(\frac{1}{k^2}\right) \leq \frac{\hat{V}}{(k+1)^2},$$

or equivalently,

$$\hat{V} \geq \frac{\theta^2\rho_w^2 c_1}{\mu^2}\left[\left(\frac{k}{k+1}\right)^2 - \frac{3+\rho_w^2}{4}\right]^{-1}. \tag{26}$$

Since $k \geq K_1 \geq \frac{16}{1-\rho_w^2}$, we have

$$\begin{aligned}
\left(\frac{k}{k+1}\right)^2 - \frac{3+\rho_w^2}{4} &= -\frac{2}{k+1} + \frac{1}{(k+1)^2} + \frac{1-\rho_w^2}{4} \\
&\geq \frac{1-\rho_w^2}{8}.
\end{aligned}$$

Hence relation (26) is satisfied with $\hat{V} \geq \frac{8\theta^2\rho_w^2 c_1}{\mu^2(1-\rho_w^2)}$. We then have for all $k \geq K_1$,

$$\tilde{V}(k) \leq \frac{1}{k^2}\max\left\{K_1^2\hat{X}, \frac{8\theta^2\rho_w^2 c_1}{\mu^2(1-\rho_w^2)}\right\}.$$

Recalling the connection of $\tilde{V}(k)$ and $V(k)$, we conclude that $V(k) \leq \frac{\hat{V}}{\tilde{k}^2}$ for all $k \geq K_1 - K$. □

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3126253, IEEE Transactions on Automatic Control

To prove the sublinear convergence of $U(k)$, we start with a useful lemma which provides lower and upper bounds for the product of a decreasing sequence. Such products arise in the convergence proof for $U(k)$ and our main convergence results in Section IV.

Lemma 11. *For any $1<a<k$ ($a\in\mathbb{N}$) and $1<\gamma\le a/2$,*

$$\frac{a^{2\gamma}}{k^{2\gamma}}\le\prod_{t=a}^{k-1}\left(1-\frac{\gamma}{t}\right)\le\frac{a^{\gamma}}{k^{\gamma}}.$$

Proof. Denote $G(k):=\prod_{t=a}^{k-1}\left(1-\frac{\gamma}{t}\right)$. We first show that $G(k)\le\frac{a^\gamma}{k^\gamma}$. Suppose $G(k)\le\frac{M_1}{k^\gamma}$ for some $M_1>0$ and $k\ge a$. Then,

$$G(k+1)=\left(1-\frac{\gamma}{k}\right)G(k)\le\left(1-\frac{\gamma}{k}\right)\frac{M_1}{k^\gamma}\le\frac{M_1}{(k+1)^\gamma}.$$

To see why the last inequality holds, note that $\left(\frac{k}{k+1}\right)^\gamma\ge 1-\frac{\gamma}{k}$. Taking $M_1=a^\gamma$, we have $G(a)=1\le\frac{M_1}{a^\gamma}$. The desired relation then holds for all $k>a$.

Now suppose $G(k)\ge\frac{M_2}{k^{2\gamma}}$ for some $M_2>0$ and $k\ge a$. It follows that

$$G(k+1)=\left(1-\frac{\gamma}{k}\right)G(k)\ge\left(1-\frac{\gamma}{k}\right)\frac{M_2}{k^{2\gamma}}\ge\frac{M_2}{(k+1)^{2\gamma}},$$

where the last inequality follows from $\left(\frac{k}{k+1}\right)^{2\gamma}\le 1-\frac{\gamma}{k}$ (noting that $\gamma\le a/2\le k/2$). Taking $M_2=a^{2\gamma}$, we have $G(a)=1\le\frac{M_2}{a^{2\gamma}}$. The desired relation then holds for all $k>a$. □

In light of Lemma 6 and the other supporting lemmas, we establish the $\mathcal{O}(\frac{1}{k})$ convergence rate of $U(k)$ in the following lemma.

Lemma 12. *Suppose Assumptions 1-3 hold. Under Algorithm* (4) *with stepsize* (12)*, suppose $\theta>2$. We have*

$$U(k)\le\frac{\theta^2c_2}{(1.5\theta-1)n\mu^2\tilde{k}}+\frac{K_1^{1.5\theta}}{\tilde{k}^{1.5\theta}}\frac{\hat{X}}{n}+\left[\frac{3\theta^2(1.5\theta-1)c_2}{(1.5\theta-2)n\mu^2}+\frac{6\theta L^2\hat{V}}{(1.5\theta-2)n\mu^2}\right]\frac{1}{\tilde{k}^2},$$

for all $k\ge K_1-K$, where

$$c_2:=\frac{2ML^2}{n}\hat{X}+\bar{M}. \tag{27}$$

Proof. In light of Lemma 6 and Lemma 8, for all $k\ge 0$, we have

$$U(k+1)\le\left(1-\frac{3}{2}\alpha_k\mu\right)U(k)+\frac{3\alpha_kL^2}{n\mu}V(k)+\frac{\alpha_k^2c_2}{n}.$$

Recalling the definitions of $\tilde{U}(k)$ and $\tilde{V}(k)$, for all $k\ge K$,

$$\tilde{U}(k+1)\le\left(1-\frac{3\theta}{2k}\right)\tilde{U}(k)+\frac{3\theta L^2}{n\mu^2}\frac{\tilde{V}(k)}{k}+\frac{\theta^2c_2}{n\mu^2}\frac{1}{k^2}.$$

Therefore,

$$\tilde{U}(k)\le\prod_{t=K_1}^{k-1}\left(1-\frac{3\theta}{2t}\right)\tilde{U}(K_1)+\sum_{t=K_1}^{k-1}\left(\prod_{j=t+1}^{k-1}\left(1-\frac{3\theta}{2j}\right)\right)\left(\frac{\theta^2c_2}{n\mu^2}\frac{1}{t^2}+\frac{3\theta L^2}{n\mu^2}\frac{\tilde{V}(t)}{t}\right).$$

From Lemma 11,

$$\begin{aligned}\tilde{U}(k)&\le\frac{K_1^{1.5\theta}}{k^{1.5\theta}}\tilde{U}(K_1)+\sum_{t=K_1}^{k-1}\frac{(t+1)^{1.5\theta}}{k^{1.5\theta}}\left(\frac{\theta^2c_2}{n\mu^2t^2}+\frac{3\theta L^2}{n\mu^2}\frac{\tilde{V}(t)}{t}\right)\\&=\frac{1}{k^{1.5\theta}}\frac{\theta^2c_2}{n\mu^2}\sum_{t=K_1}^{k-1}\frac{(t+1)^{1.5\theta}}{t^2}+\frac{K_1^{1.5\theta}}{k^{1.5\theta}}\tilde{U}(K_1)\\&\quad+\sum_{t=K_1}^{k-1}\frac{(t+1)^{1.5\theta}}{k^{1.5\theta}}\frac{3\theta L^2}{n\mu^2}\frac{\tilde{V}(t)}{t}.\end{aligned}$$

In light of Lemma 10, when $k\ge K_1$, $\tilde{V}(k)\le\frac{\hat{V}}{k^2}$. Hence,

$$\begin{aligned}&\tilde{U}(k)-\frac{1}{k^{1.5\theta}}\frac{\theta^2c_2}{n\mu^2}\sum_{t=K_1}^{k-1}\frac{(t+1)^{1.5\theta}}{t^2}-\frac{K_1^{1.5\theta}}{k^{1.5\theta}}\tilde{U}(K_1)\\&\le\sum_{t=K_1}^{k-1}\frac{(t+1)^{1.5\theta}}{k^{1.5\theta}}\frac{3\theta L^2}{n\mu^2}\frac{\hat{V}}{t^3}\\&=\frac{1}{k^{1.5\theta}}\frac{3\theta L^2\hat{V}}{n\mu^2}\sum_{t=K_1}^{k-1}\frac{(t+1)^{1.5\theta}}{t^3}.\end{aligned}$$

However, we have for any $b>a\ge K_1$,

$$\begin{aligned}&\sum_a^b\frac{(t+1)^{1.5\theta}}{t^2}\\&\le\sum_a^{b-2}\left[\frac{(t+1)^{1.5\theta}}{(t+1)^2}+3\frac{(t+1)^{1.5\theta}}{(t+1)^3}\right]+\frac{b^{1.5\theta}}{(b-1)^2}\\&\quad+\frac{(b+1)^{1.5\theta}}{b^2}\\&\le\int_a^b\left(t^{1.5\theta-2}+3t^{1.5\theta-3}\right)dt+\frac{2(b+1)^{1.5\theta}}{b^2}\\&\le\frac{b^{1.5\theta-1}}{1.5\theta-1}+\frac{3b^{1.5\theta-2}}{1.5\theta-2}+3b^{1.5\theta-2},\end{aligned}$$

where the last inequality comes from the fact that $(\frac{b+1}{b})^{1.5\theta}\le(\frac{4\theta+1}{4\theta})^{1.5\theta}\le\exp(\frac{3}{8})<\frac{3}{2}$ (given that $b>K_1\ge 4\theta$), and

$$\begin{aligned}&\sum_a^b\frac{(t+1)^{1.5\theta}}{t^3}\le\frac{3}{2}\sum_a^b t^{1.5\theta-3}\le\frac{3}{2}\int_a^{b+1}t^{1.5\theta-3}dt\\&\le\frac{3}{2}\frac{(b+1)^{1.5\theta-2}}{1.5\theta-2}\le\frac{2b^{1.5\theta-2}}{1.5\theta-2}.\end{aligned}$$

Hence, for all $k\ge K_1$,

$$\begin{aligned}\tilde{U}(k)\le{}&\frac{\theta^2c_2}{(1.5\theta-1)n\mu^2k}+\frac{3\theta^2(1.5\theta-1)c_2}{(1.5\theta-2)n\mu^2}\frac{1}{k^2}\\&+\frac{K_1^{1.5\theta}}{k^{1.5\theta}}\tilde{U}(K_1)+\frac{6\theta L^2\hat{V}}{(1.5\theta-2)n\mu^2}\frac{1}{k^2}.\end{aligned}$$

Recalling Lemma 9 and the definition of $\tilde{U}(k)$ yields the desired result. □

Remark 1. Notice that the convergence rate established in Lemma 12 is not asymptotically the same as centralized stochastic gradient descent, since the constant c_2 contains information about the initial solutions. In the next section, we will improve the convergence result and show that DSGD indeed performs as well as centralized SGD asymptotically.

IV. Main Results

In this section, we perform a non-asymptotic analysis of network independence for Algorithm (4). Specifically, in Theorem 1, we show that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|x_i(k)-x_*\|^2] = \frac{\theta^2\bar{M}}{(2\theta-1)n\mu^2\tilde{k}} + \mathcal{O}\left(\frac{1}{\sqrt{n}(1-\rho_w)}\right)\frac{1}{\tilde{k}^{1.5}} + \mathcal{O}\left(\frac{1}{(1-\rho_w)^2}\right)\frac{1}{\tilde{k}^2},$$

where the first term is network independent and the second and third (higher-order) terms depends on $(1-\rho_w)$. Then we compare the result with centralized stochastic gradient descent and show that asymptotically, the two methods have the same convergence rate $\frac{\theta^2\bar{M}}{(2\theta-1)n\mu^2\tilde{k}}$. In addition, it takes $K_T = \mathcal{O}\left(\frac{n}{(1-\rho_w)^2}\right)$ time for Algorithm (4) to reach this asymptotic rate of convergence. Finally, we construct a "hard" optimization problem for which we show the transient time K_T is sharp.

Our first step is to simplify the presentation of the convergence results in Lemma 10 and Lemma 12, so that we can utilize them for deriving improved convergence rates conveniently. For this purpose, we first estimate the constants $\hat{X}$, $\hat{V}$, c_1 and c_2 appearing in the two lemmas and derive their dependency on the network size n, the spectral gap $(1-\rho_w)$, the summation of initial optimization errors $\sum_{i=1}^n\|x_i(0)-x_*\|^2$, and $\sum_{i=1}^n\|\nabla f_i(x_*)\|^2$, where the last term can be seen as a measure of the difference among each agent's individual cost functions.

Lemma 13. *Denote* $A := \sum_{i=1}^n\|x_i(0)-x_*\|^2$ *and* $B := \sum_{i=1}^n\|\nabla f_i(x_*)\|^2$. *Then,*

$$\hat{X} = \mathcal{O}(A+B+n), \quad c_1 = \mathcal{O}\left(\frac{A+B+n}{1-\rho_w}\right),$$
$$\hat{V} = \mathcal{O}\left(\frac{A+B+n}{(1-\rho_w)^2}\right), \quad c_2 = \mathcal{O}\left(\frac{A+B+n}{n}\right).$$

Proof. We first estimate the constant $\hat{X}$ which appears in the definition (23) for $\hat{V}$. From Lemma 8,

$$\hat{X} \le \|\mathbf{x}(0)-\mathbf{1}x_*\|^2 + \frac{9\|\nabla F(\mathbf{1}x_*^{\intercal})\|^2}{\mu^2} + \frac{n\sigma^2}{(1+M)L^2} = \mathcal{O}(A+B+n).$$

From the definition of c_1 in (24),

$$c_1 = 2\left(\frac{3}{1-\rho_w^2}+M\right)\left(L^2\hat{X}+\|\nabla F(\mathbf{1}x_*^{\intercal})\|^2\right)+n\sigma^2 = \mathcal{O}\left(\frac{A+B+n}{1-\rho_w}\right).$$

Noting that $K_1 = \mathcal{O}(\frac{1}{1-\rho_w})$, by definition,

$$\hat{V} = \max\left\{K_1^2\hat{X}, \frac{8\theta^2\rho_w^2c_1}{\mu^2(1-\rho_w^2)}\right\} = \mathcal{O}\left(\frac{A+B+n}{(1-\rho_w)^2}\right).$$

From the definition of c_2 in (27), we have

$$c_2 = \frac{2ML^2}{n}\hat{X}+\bar{M} = \mathcal{O}\left(\frac{A+B+n}{n}\right).$$

□

In light of Lemma 13, the convergence result of $V(k)$ given in Lemma 10 can be easily simplified since $\hat{V}$ is the only constant. Regarding the optimization error $U(k)$, in light of Lemma 8, Lemma 10, Lemma 12 and Lemma 13, we have the following corollary which simplifies the presentation of the convergence result in Lemma 12.

Corollary 1. *Suppose Assumptions 1-3 hold. Under Algorithm* (4) *with stepsize* (12) *and assuming* $\theta > 2$, *when* $k \ge K_1 - K$,

$$U(k) \le \frac{\theta^2c_2}{(1.5\theta-1)n\mu^2\tilde{k}} + \frac{c}{\tilde{k}^2},$$

where

$$c = \mathcal{O}\left(\frac{A+B+n}{n(1-\rho_w)^2}\right).$$

Proof. From Theorem 12 and Lemma 13, when $k \ge K_1 - K = \mathcal{O}(\frac{1}{1-\rho_w})$,

$$\begin{aligned} U(k) \le& \frac{\theta^2c_2}{(1.5\theta-1)n\mu^2\tilde{k}} + \frac{K_1^{1.5\theta-2}}{\tilde{k}^{1.5\theta-2}}\frac{\hat{X}}{n}\frac{K_1^2}{\tilde{k}^2} \\ &+ \left[\frac{3\theta^2(1.5\theta-1)c_2}{(1.5\theta-2)n\mu^2} + \frac{6\theta L^2\hat{V}}{(1.5\theta-2)n\mu^2}\right]\frac{1}{\tilde{k}^2} \\ =& \frac{\theta^2c_2}{(1.5\theta-1)n\mu^2}\frac{1}{\tilde{k}} + \mathcal{O}\left(\frac{A+B+n}{n(1-\rho_w)^2}\right)\frac{1}{\tilde{k}^2}. \end{aligned} \tag{28}$$

□

Let $\frac{1}{n}\sum_{i=1}^n\mathbb{E}[\|x_i(k)-x_*\|^2]$, the average optimization error for each agent to measure the performance of DSGD. In the following theorem, we improve the result of Corollary 1 with further analysis and derive the main convergence result for Algorithm (4).

Theorem 1. *Suppose Assumptions 1-3 hold. Under Algorithm* (4) *with stepsize* (12) *and assuming* $\theta > 2$, *when* $k \ge K_1 - K$,

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|x_i(k)-x_*\|^2] \le& \frac{\theta^2\bar{M}}{(2\theta-1)n\mu^2\tilde{k}} \\ &+ \mathcal{O}\left(\frac{\sqrt{A+B+n}}{n(1-\rho_w)}\right)\frac{1}{\tilde{k}^{1.5}} + \mathcal{O}\left(\frac{A+B+n}{n(1-\rho_w)^2}\right)\frac{1}{\tilde{k}^2}. \end{aligned} \tag{29}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3126253, IEEE Transactions on Automatic Control

Proof. For $k \geq K_1 - K$, in light of Lemma 2 and Lemma 5,

$$\begin{aligned}
&U(k+1) \\
&\leq (1-\alpha_k\mu)^2 U(k) + \frac{2\alpha_k L}{\sqrt{n}}\mathbb{E}[\|\overline{x}(k)-x_*\|\|\mathbf{x}(k)-\mathbf{1}\overline{x}(k)^\intercal\|] \\
&\quad + \frac{\alpha_k^2 L^2}{n}V(k) + \alpha_k^2\left(\frac{2ML^2}{n^2}\mathbb{E}[\|\mathbf{x}(k)-\mathbf{1}x_*^\intercal\|^2] + \frac{\bar{M}}{n}\right) \\
&\leq (1-\alpha_k\mu)^2 U(k) + \frac{2\alpha_k L}{\sqrt{n}}\sqrt{U(k)V(k)} + \frac{\alpha_k^2 L^2}{n}V(k) \\
&\quad + \alpha_k^2\left[\frac{2ML^2}{n^2}(nU(k)+V(k)) + \frac{\bar{M}}{n}\right] \\
&= (1-2\alpha_k\mu)U(k) + \alpha_k^2\left(\mu^2 + \frac{2ML^2}{n}\right)U(k) \\
&\quad + \frac{2\alpha_k L}{\sqrt{n}}\sqrt{U(k)V(k)} + \frac{\alpha_k^2 L^2}{n}\left(1+\frac{2M}{n}\right)V(k) + \frac{\alpha_k^2\bar{M}}{n},
\end{aligned}$$

where the second inequality follows from the Cauchy-Schwarz inequality (see [52], page 62) and the fact that $\|\mathbf{x}(k) - \mathbf{1}x_*^\intercal\|^2 = \|\mathbf{x}(k) - \mathbf{1}\overline{x}(k)\|^2 + n\|\overline{x}(k) - x_*\|^2$.

Recalling the definitions of $\tilde{U}(k)$ and $\tilde{V}(k)$, when $k \geq K_1$,

$$\begin{aligned}
\tilde{U}(k+1) &\leq \left(1-\frac{2\theta}{k}\right)\tilde{U}(k) + \frac{\theta^2}{k^2}\left(1+\frac{2ML^2}{n\mu^2}\right)\tilde{U}(k) \\
&+ \frac{2\theta L}{\sqrt{n}\mu}\frac{\sqrt{\tilde{U}(k)\tilde{V}(k)}}{k} + \frac{\theta^2 L^2}{n\mu^2}\left(1+\frac{2M}{n}\right)\frac{\tilde{V}(k)}{k^2} + \frac{\theta^2\bar{M}}{n\mu^2}\frac{1}{k^2}.
\end{aligned}$$

Therefore, by denoting $c_3 := 1 + \frac{2ML^2}{n\mu^2}$ and $c_4 := 1 + \frac{2M}{n}$, we have

$$\begin{aligned}
&\tilde{U}(k) \leq \left(\prod_{t=K_1}^{k-1}\left(1-\frac{2\theta}{t}\right)\right)\tilde{U}(K_1) \\
&\quad + \sum_{t=K_1}^{k-1}\left(\prod_{i=t+1}^{k-1}\left(1-\frac{2\theta}{i}\right)\right) \\
&\quad \cdot \left(\frac{\theta^2\bar{M}}{n\mu^2 t^2} + \frac{\theta^2 c_3\tilde{U}(t)}{t^2} + \frac{2\theta L}{\sqrt{n}\mu}\frac{\sqrt{\tilde{U}(t)\tilde{V}(t)}}{t} + \frac{\theta^2 L^2 c_4}{n\mu^2}\frac{\tilde{V}(t)}{t^2}\right)
\end{aligned}$$

From Lemma 11,

$$\begin{aligned}
\tilde{U}(k) &\leq \frac{K_1^{2\theta}}{k^{2\theta}}\tilde{U}(K_1) + \sum_{t=K_1}^{k-1}\frac{(t+1)^{2\theta}}{k^{2\theta}}\left(\frac{\theta^2\bar{M}}{n\mu^2 t^2} + \frac{\theta^2 c_3\tilde{U}(t)}{t^2}\right. \\
&\quad \left. + \frac{2\theta L}{\sqrt{n}\mu}\frac{\sqrt{\tilde{U}(t)\tilde{V}(t)}}{t} + \frac{\theta^2 L^2 c_4}{n\mu^2}\frac{\tilde{V}(t)}{t^2}\right) \\
&= \frac{1}{k^{2\theta}}\frac{\theta^2\bar{M}}{n\mu^2}\sum_{t=K_1}^{k-1}\frac{(t+1)^{2\theta}}{t^2} + \frac{K_1^{2\theta}}{k^{2\theta}}\tilde{U}(K_1) \\
&\quad + \sum_{t=K_1}^{k-1}\frac{(t+1)^{2\theta}}{k^{2\theta}}\left(\frac{\theta^2 c_3\tilde{U}(t)}{t^2} + \frac{2\theta L}{\sqrt{n}\mu}\frac{\sqrt{\tilde{U}(t)\tilde{V}(t)}}{t}\right. \\
&\quad \left. + \frac{\theta^2 L^2 c_4}{n\mu^2}\frac{\tilde{V}(t)}{t^2}\right).
\end{aligned}$$

Hence, by Corollary 1,

$$\begin{aligned}
&\tilde{U}(k) - \frac{1}{k^{2\theta}}\frac{\theta^2\bar{M}}{n\mu^2}\sum_{t=K_1}^{k-1}\frac{(t+1)^{2\theta}}{t^2} - \frac{K_1^{2\theta}}{k^{2\theta}}\tilde{U}(K_1) \\
&\leq \frac{\theta^2 c_3}{k^{2\theta}}\sum_{t=K_1}^{k-1}\frac{(t+1)^{2\theta}}{t^2}\left[\frac{\theta^2 c_2}{(1.5\theta-1)n\mu^2 t} + \frac{c}{t^2}\right] \\
&\quad + \frac{1}{k^{2\theta}}\frac{2\theta L}{\sqrt{n}\mu}\sum_{t=K_1}^{k-1}\frac{(t+1)^{2\theta}}{t}\sqrt{\frac{\theta^2 c_2}{(1.5\theta-1)n\mu^2}\frac{1}{t} + \frac{c}{t^2}}\sqrt{\frac{\hat{V}}{t^2}} \\
&\quad + \frac{1}{k^{2\theta}}\frac{\theta^2 L^2 c_4}{n\mu^2}\sum_{t=K_1}^{k-1}\frac{(t+1)^{2\theta}}{t^2}\frac{\hat{V}}{t^2}.
\end{aligned}$$

Since

$$\sqrt{\frac{\theta^2 c_2}{(1.5\theta-1)n\mu^2}\frac{1}{t} + \frac{c}{t^2}}\sqrt{\frac{\hat{V}}{t^2}} \leq \sqrt{\frac{\theta^2 c_2\hat{V}}{(1.5\theta-1)n\mu^2}\frac{1}{t^3}} + \frac{\sqrt{c\hat{V}}}{t^2},$$

we have

$$\begin{aligned}
&\tilde{U}(k) - \frac{1}{k^{2\theta}}\frac{\theta^2\bar{M}}{n\mu^2}\sum_{t=K_1}^{k-1}\frac{(t+1)^{2\theta}}{t^2} - \frac{K_1^{2\theta}}{k^{2\theta}}\tilde{U}(K_1) \\
&\leq \frac{\theta^2 c_3}{k^{2\theta}}\sum_{t=K_1}^{k-1}\frac{(t+1)^{2\theta}}{t^2}\left[\frac{\theta^2 c_2}{(1.5\theta-1)n\mu^2}\frac{1}{t} + \frac{c}{t^2}\right] \\
&\quad + \frac{1}{k^{2\theta}}\frac{2\theta L}{\sqrt{n}\mu}\sum_{t=K_1}^{k-1}\frac{(t+1)^{2\theta}}{t}\left(\sqrt{\frac{\theta^2 c_2\hat{V}}{(1.5\theta-1)n\mu^2}\frac{1}{t^3}} + \frac{\sqrt{c\hat{V}}}{t^2}\right) \\
&\quad + \frac{1}{k^{2\theta}}\frac{\theta^2 L^2 c_4\hat{V}}{n\mu^2}\sum_{t=K_1}^{k-1}\frac{(t+1)^{2\theta}}{t^4} \\
&= \frac{1}{k^{2\theta}}\left(\frac{2\theta^2 L\sigma\sqrt{c_2\hat{V}}}{\sqrt{1.5\theta-1}n\mu^2}\right)\sum_{t=K_1}^{k-1}\frac{(t+1)^{2\theta}}{t^{2.5}} \\
&\quad + \frac{1}{k^{2\theta}}\left(\frac{\theta^4 c_3 c_2}{(1.5\theta-1)n\mu^2} + \frac{2\theta L\sqrt{c\hat{V}}}{\sqrt{n}\mu}\right)\sum_{t=K_1}^{k-1}\frac{(t+1)^{2\theta}}{t^3} \\
&\quad + \frac{1}{k^{2\theta}}\left(\theta^2 c_3 c + \frac{\theta^2 L^2 c_4\hat{V}}{n\mu^2}\right)\sum_{t=K_1}^{k-1}\frac{(t+1)^{2\theta}}{t^4}.
\end{aligned}$$

Notice that $c_2 = \mathcal{O}(\frac{A+B+n}{n})$ and $c_3, c_4 = \mathcal{O}(1)$. Following a discussion similar to those in the proofs for Theorem 12 and Corollary 1, we have

$$\begin{aligned}
\tilde{U}(k) \leq& \frac{\theta^2\bar{M}}{(2\theta-1)n\mu^2 k} + \mathcal{O}\left(\frac{\sqrt{A+B+n}}{n(1-\rho_w)}\right)\frac{1}{k^{1.5}} \\
&+ \mathcal{O}\left(\frac{A+B+n}{n(1-\rho_w)^2}\right)\frac{1}{k^2} + \mathcal{O}\left(\frac{A+B+n}{n(1-\rho_w)^2}\right)\frac{1}{k^3} \\
&+ \mathcal{O}\left(\frac{A+B+n}{n(1-\rho_w)^{2\theta}}\right)\frac{1}{k^{2\theta}} \\
=& \frac{\theta^2\bar{M}}{(2\theta-1)n\mu^2 k} + \mathcal{O}\left(\frac{\sqrt{A+B+n}}{n(1-\rho_w)}\right)\frac{1}{k^{1.5}} \\
&+ \mathcal{O}\left(\frac{A+B+n}{n(1-\rho_w)^2}\right)\frac{1}{k^2}.
\end{aligned}$$

Noting that

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|x_i(k)-x_*\|^2]\\&=\mathbb{E}[\|\bar{x}(k)-x_*\|^2]+\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|x_i(k)-\bar{x}\|^2]\\&=U(k)+\frac{V(k)}{n},\end{aligned}$$

and $U(k)=\tilde{U}(k+K)$, in light of the bound on $V(k)$ in Lemma 10 and the estimate of $\hat{V}$ in Lemma 13, we obtain the desired result. □

A. Comparison with Centralized Implementation

We compare the performance of DSGD and centralized stochastic gradient descent (SGD) stated below:

$$x(k+1)=x(k)-\alpha_k\tilde{g}(k), \tag{30}$$

where $\alpha_k:=\frac{\theta}{\mu k}$ $(\theta>1)$ and $\tilde{g}(k):=\frac{1}{n}\sum_{i=1}^{n}g(x(k),\xi_i(k))$.

First, we derive the convergence rate for SGD which matches the optimal rate for such stochastic gradient methods (see [29], [53]). Our result relies on an analysis different from the literature that considered a compact feasible set and uniformly bounded stochastic gradients in expectation.

Theorem 2. *Under the centralized stochastic gradient descent of (30), suppose* $k\geq K_2:=\left\lceil\frac{\theta L}{\mu}\right\rceil$. *We have*

$$\mathbb{E}[\|x(k)-x_*\|^2]\leq\frac{\theta^2\bar{M}}{(2\theta-1)n\mu^2 k}+\mathcal{O}\left(\frac{1}{n}\right)\frac{1}{k^2}.$$

Proof. Noting that $\alpha_k\leq 1/L$ when $k\geq K_2$, we have from Lemma 3 that

$$\begin{aligned}&\mathbb{E}[\|x(k+1)-x_*\|^2\mid\mathcal{F}(k)]\\&=\mathbb{E}[\|x(k)-\alpha_k\tilde{g}(k)-x_*\|^2\mid\mathcal{F}(k)]\\&=\|x(k)-\alpha_k\nabla f(x(k))-x_*\|^2+\alpha_k^2\mathbb{E}[\|\nabla f(x(k))-\tilde{g}(k)\|^2]\\&\leq(1-\alpha_k\mu)^2\|x(k)-x_*\|^2\\&\quad+\alpha_k^2\left(\frac{2ML^2}{n}\|x(k)-x_*\|^2+\frac{\bar{M}}{n}\right)\\&=\left(1-\frac{2\theta}{k}\right)\|x(k)-x_*\|^2\\&\quad+\theta^2\left(1+\frac{2ML^2}{n\mu^2}\right)\|x(k)-x_*\|^2\frac{1}{k^2}+\frac{\theta^2\bar{M}}{n\mu^2}\frac{1}{k^2}.\end{aligned} \tag{31}$$

It can be shown first that $\mathbb{E}[\|x(k)-x_*\|^2]\leq\frac{c_5}{k}$ for $k\geq K_2$, where $c_5=\mathcal{O}(\frac{1}{n})$.[3] Denote $\bar{c}_5:=(1+\frac{2ML^2}{n\mu^2})c_5$. Then from relation (31), when $k\geq K_2$,

$$\begin{aligned}\mathbb{E}[\|x(k)-x_*\|^2]\leq&\left(\prod_{t=K_2}^{k-1}\left(1-\frac{2\theta}{t}\right)\right)\mathbb{E}[\|x(K_2)-x_*\|^2]\\&+\sum_{t=K_2}^{k-1}\left(\prod_{i=t+1}^{k-1}\left(1-\frac{2\theta}{i}\right)\right)\left(\frac{\theta^2\bar{M}}{n\mu^2t^2}+\frac{\theta^2\bar{c}_5}{t^3}\right).\end{aligned}$$

[3] The argument here is similar to that in the proof for Lemma 12.

From Lemma 11,

$$\begin{aligned}&\mathbb{E}[\|x(k)-x_*\|^2]\leq\frac{K_2^{2\theta}}{k^{2\theta}}\mathbb{E}[\|x(K_2)-x_*\|^2]\\&\quad+\sum_{t=K_2}^{k-1}\frac{(t+1)^{2\theta}}{k^{2\theta}}\left(\frac{\theta^2\bar{M}}{n\mu^2t^2}+\frac{\theta^2\bar{c}_5}{t^3}\right)\\&=\frac{1}{k^{2\theta}}\frac{\theta^2\bar{M}}{n\mu^2}\sum_{t=K_2}^{k-1}\frac{(t+1)^{2\theta}}{t^2}+\frac{K_2^{2\theta}}{k^{2\theta}}\mathbb{E}[\|x(K_2)-x_*\|^2]\\&\quad+\frac{\theta^2\bar{c}_5}{k^{2\theta}}\sum_{t=K_2}^{k-1}\frac{(t+1)^{2\theta}}{t^3}\\&=\frac{\theta^2\bar{M}}{(2\theta-1)n\mu^2k}+\mathcal{O}\left(\frac{1}{n}\right)\frac{1}{k^2}.\end{aligned}$$

□

Comparing the results of Theorem 1 and Theorem 2, we can see that asymptotically, DSGD and SGD have the same convergence rate $\frac{\theta^2\bar{M}}{(2\theta-1)n\mu^2k}$. The next corollary identifies the time needed for DSGD to achieve this rate.

Corollary 2 (Transient Time)**.** *Suppose Assumptions 1-3 hold. Assume in addition that* $\sum_{i=1}^{n}\|x_i(0)-x_*\|^2=\mathcal{O}(n)$ *and* $\sum_{i=1}^{n}\|\nabla f_i(x_*)\|^2=\mathcal{O}(n)$. *It takes* $K_T=\mathcal{O}\left(\frac{n}{(1-\rho_w)^2}\right)$ *time for Algorithm* (4) *to reach the asymptotic rate of convergence, i.e., when* $k\geq K_T$, *we have* $\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|x_i(k)-x_*\|^2]\leq\frac{\theta^2\bar{M}}{(2\theta-1)n\mu^2k}\mathcal{O}(1)$.

Proof. From (29),

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|x_i(k)-x_*\|^2]\leq&\frac{\theta^2\bar{M}}{(2\theta-1)n\mu^2k}\\&\cdot\left[1+\mathcal{O}\left(\frac{\sqrt{n}}{(1-\rho_w)}\right)\frac{1}{k^{0.5}}+\mathcal{O}\left(\frac{n}{(1-\rho_w)^2}\right)\frac{1}{k}\right].\end{aligned}$$

Let K_T be such that

$$\mathcal{O}\left(\frac{\sqrt{n}}{(1-\rho_w)}\right)\frac{1}{K_T^{0.5}}+\mathcal{O}\left(\frac{n}{(1-\rho_w)^2}\right)\frac{1}{K_T}=\mathcal{O}(1).$$

We then obtain that

$$K_T=\mathcal{O}\left(\frac{n}{(1-\rho_w)^2}\right).$$

□

Remark 2. By assuming the additional conditions $\sum_{i=1}^{n}\|x_i(0)-x_*\|^2=\mathcal{O}(n)$ and $\sum_{i=1}^{n}\|\nabla f_i(x_*)\|^2=\mathcal{O}(n)$, motivated by the observation that each of these expression is the sum of n terms, we obtain a cleaner expression of the transient time. In general, we would obtain $K_T=\mathcal{O}\left(\frac{A+B+n}{(1-\rho_w)^2}\right)$.

Remark 3. For general connected networks such as line graphs, if we adopt the Lazy Metropolis rule for choosing the weights $[w_{ij}]$ (see [54]), then $\frac{1}{1-\rho_w}=\mathcal{O}(n^2)$, and hence $K_T=\mathcal{O}(n^5)$. The transient time can be improved for networks with special structures. For example, $\frac{1}{1-\rho_w}$ is constant with high probability for a random Erdős-Rényi random graph, and consequently $K_T=\mathcal{O}(n)$ on such a graph.

The next theorem states that the transient time for DSGD to reach the asymptotic convergence rate is lower bounded by $\Omega\left(\frac{n}{(1-\rho_w)^2}\right)$, that is, under Assumptions 1-3 and assuming $\sum_{i=1}^n \|x_i(0)-x_*\|^2 = \mathcal{O}(n)$ and $\sum_{i=1}^n \|\nabla f_i(x_*)\|^2 = \mathcal{O}(n)$, there exists an optimization problem whose transient time under DSGD is lower bounded by $\Omega\left(\frac{n}{(1-\rho_w)^2}\right)$. This implies the result in Corollary 2 is sharp and can not be improved in general.

Theorem 3. *Suppose Assumptions 1-3 hold. Assume in addition that* $\sum_{i=1}^n \|x_i(0) - x_*\|^2 = \mathcal{O}(n)$ *and* $\sum_{i=1}^n \|\nabla f_i(x_*)\|^2 = \mathcal{O}(n)$. *Then there exists a* $\rho_0 \in (0,1)$ *such that if* $\rho_w \geq \rho_0$, *then the time needed for DSGD to reach the asymptotic convergence rate is lower bounded by* $\Omega\left(\frac{n}{(1-\rho_w)^2}\right)$.

Proof. We construct a "hard" optimization to prove the claimed result, inspired by [31]. Consider quadratic objective functions $f_i(x) := \frac{1}{2}\|x - x_i^*\|^2$, where $x, x_i^* \in \mathbb{R}$. The optimal solution to Problem (1) is given by $x_* = \frac{1}{n}\sum_{i=1}^n x_i^*$. The DSGD algorithm implements:

$$\mathbf{x}(k+1) = \mathbf{W}\left(\mathbf{x}(k) - \alpha_k(\mathbf{x}(k) - \mathbf{x}_*) + \alpha_k \mathbf{n}(k)\right), \quad (32)$$

where $\mathbf{x}_* := [x_1^*, x_2^*, \ldots, x_n^*]^\intercal$, and $\mathbf{n}(k)$ denotes the vector of gradient noise terms. From (12), we use stepsize $\alpha_k = \frac{\theta}{k+K}$ ($\theta > 2$), where $K = \lceil 2\theta \rceil$ since $\mu = L = 1$. We rewrite (32) as

$$\mathbf{x}(k+1) = (1-\alpha_k)\mathbf{W}\mathbf{x}(k) + \alpha_k \mathbf{W}\mathbf{x}_* + \alpha_k \mathbf{W}\mathbf{n}(k).$$

It follows that

$$\begin{aligned}\mathbf{x}(k+1) - \mathbf{1}\overline{x}(k+1)^\intercal &= (1-\alpha_k)\mathbf{W}(\mathbf{x}(k) - \mathbf{1}\overline{x}(k)^\intercal) \\ &\quad + \alpha_k \mathbf{W}(\mathbf{x}_* - \mathbf{1}x_*) + \alpha_k \mathbf{W}(\mathbf{n}(k) - \mathbf{1}\overline{n}(k)).\end{aligned}$$

By induction, we have for all $k > 0$,

$$\begin{aligned}\mathbf{x}(k) - \mathbf{1}\overline{x}(k)^\intercal &= \left(\prod_{t=0}^{k-1}(1-\alpha_t)\right)\mathbf{W}^k(\mathbf{x}(0) - \mathbf{1}\overline{x}(0)) \\ &\quad + \sum_{t=0}^{k-1}\left[\left(\prod_{j=t+1}^{k-1}(1-\alpha_j)\right)\alpha_t \mathbf{W}^{k-t}\right] \\ &\quad \cdot \left[(\mathbf{x}_* - \mathbf{1}x_*) + (\mathbf{n}(t) - \mathbf{1}\overline{n}(t))\right]. \end{aligned} \quad (33)$$

Assume that: (i) the matrix $\mathbf{W}$ is symmetric; (ii) $\mathbf{W}\mathbf{x}_* = \rho_w \mathbf{x}_*$, i.e., $\mathbf{x}_*$ is an eigenvector of $\mathbf{W}$ w.r.t. eigenvalue ρ_w (hence $x_* = \frac{1}{n}\mathbf{1}^\intercal \mathbf{x}_* = 0$); (iii) $\|\nabla F(\mathbf{1}x_*^\intercal)\|^2 = \|\mathbf{1}x_* - \mathbf{x}_*\|^2 = \|\mathbf{x}_*\|^2 = \Omega(n)$; (iv) $\mathbf{x}(0) = \mathbf{x}_*$.[4] Then $\overline{x}(0) = x_* = 0$, and from relation (33) it follows

$$\begin{aligned}\mathbf{x}(k) - \mathbf{1}\overline{x}(k)^\intercal &= \left(\prod_{t=0}^{k-1}(1-\alpha_t)\right)\rho_w^k \mathbf{x}_* \\ &+ \sum_{t=0}^{k-1}\left[\left(\prod_{j=t+1}^{k-1}(1-\alpha_j)\right)\alpha_t \rho_w^{k-t}\right]\mathbf{x}_* + \epsilon(k), \end{aligned} \quad (34)$$

[4]Assumptions (iii) and (iv) correspond to the conditions $\|\mathbf{x}(0) - \mathbf{1}x_*^\intercal\|^2 = \mathcal{O}(n)$ and $\|\nabla F(\mathbf{1}x_*^\intercal)\|^2 = \mathcal{O}(n)$ assumed in the main results such as Theorem 1 and Corollary 2.

where $\epsilon(k)$ captures the random perturbation caused by gradient noise that has mean zero. Therefore,

$$\mathbb{E}[\|\mathbf{x}(k) - \mathbf{1}\overline{x}(k)^\intercal\|^2] \geq \left\|\sum_{t=0}^{k-1}\left[\left(\prod_{j=t+1}^{k-1}(1-\alpha_j)\right)\alpha_t \rho_w^{k-t}\right]\mathbf{x}_*\right\|^2.$$

Recalling the definition $V(k) = \mathbb{E}[\|\mathbf{x}(k) - \mathbf{1}\overline{x}(k)^\intercal\|^2$ and $\tilde{V}(k) = V(k-K)$, and noticing that $\alpha_k = \frac{\theta}{k+K}$, we have

$$\begin{aligned}\tilde{V}(k) &\geq \left\{\sum_{t=K}^{k-1}\left[\left(\prod_{j=t+1}^{k-1}\left(1-\frac{\theta}{j}\right)\right)\frac{\theta}{t}\rho_w^{k-t}\right]\right\}^2 \|\mathbf{x}_*\|^2 \\ &\geq \left\{\sum_{t=K}^{k-1}\left[\frac{(t+1)^{2\theta}}{k^{2\theta}}\frac{\theta}{t}\rho_w^{k-t}\right]\right\}^2 \|\mathbf{x}_*\|^2,\end{aligned}$$

where we invoked Lemma 11 for the second inequality. Then,

$$\begin{aligned}\tilde{V}(k) &\geq \left[\frac{\theta\rho_w^k}{k^{2\theta}}\sum_{t=K}^{k-1}\frac{(t+1)^{2\theta-1}}{\rho_w^t}\right]^2 \|\mathbf{x}_*\|^2 \\ &\geq \left[\frac{\theta\rho_w^k}{k^{2\theta}}\int_{t=K-1}^{k-1}\frac{(t+1)^{2\theta-1}}{\rho_w^t}dt\right]^2 \|\mathbf{x}_*\|^2. \end{aligned} \quad (35)$$

Note that when $k \geq \frac{4\theta}{(-\ln\rho_w)}$,

$$\begin{aligned}&\int_{t=K-1}^{k-1}\frac{(t+1)^{2\theta-1}}{\rho_w^t}dt \geq \frac{2}{3}\left.\frac{(t+1)^{2\theta-1}}{(-\ln\rho_w)\rho_w^t}\right|_{t=K-1}^{k-1} \\ &= \frac{2k^{2\theta-1}}{3(-\ln\rho_w)\rho_w^{k-1}} - \frac{2K^{2\theta-1}}{3(-\ln\rho_w)\rho_w^{K-1}} \geq \frac{k^{2\theta-1}}{2(-\ln\rho_w)\rho_w^{k-1}}.\end{aligned}$$

From (35),

$$\tilde{V}(k) \geq \left[\frac{\theta\rho_w}{2(-\ln\rho_w)k}\right]^2 \|\mathbf{x}_*\|^2 = \Omega\left(\frac{n}{(1-\rho_w)^2}\right)\frac{1}{k^2},$$

where the equality is obtained from the Taylor expansion of $\ln\rho_w$ when $\rho_w \to 1$. Since

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|x_i(k) - x_*\|^2] = U(k) + \frac{V(k)}{n} \geq \Omega\left(\frac{1}{(1-\rho_w)^2}\right)\frac{1}{k^2},$$

setting this to be at most $\frac{\bar{M}}{nk}$, we obtain that the transient time for DSGD to reach the asymptotic convergence rate is lower bounded by $\Omega\left(\frac{n}{(1-\rho_w)^2}\right)$, based on an argument similar to that of Corollary 2. □

V. Numerical Examples

In this section, we provide two numerical example to verify and complement our theoretical findings.

A. Ridge Regression

Consider the *on-line* ridge regression problem, i.e.,

$$\min_{x\in\mathbb{R}^p} f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x)\left(= \mathbb{E}_{u_i,v_i}\left[(u_i^\intercal x - v_i)^2 + \rho\|x\|^2\right]\right), \quad (36)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2021.3126253, IEEE Transactions on Automatic Control

where $\rho > 0$ is a penalty parameter. Each agent i collects data samples in the form of (u_i, v_i) continuously where $u_i \in \mathbb{R}^p$ represent the features and $v_i \in \mathbb{R}$ are the observed outputs. Assume each $u_i \in [-0.5, 0.5]^p$ is uniformly distributed, and v_i is drawn according to $v_i = u_i^\intercal \tilde{x}_i + \varepsilon_i$, where $\tilde{x}_i$ are predefined parameters evenly located in $[0, 10]^p$, and ε_i are independent Gaussian random variables (noise) with mean 0 and variance 0.01. Given a pair (u_i, v_i), agent i can compute an estimated (unbiased) gradient of $f_i(x)$: $g_i(x, u_i, v_i) = 2(u_i^\intercal x - v_i)u_i + 2\rho x$. Problem (36) has a unique solution x_* given by

$$x_* = \left(\sum_{i=1}^n \mathbb{E}_{u_i}[u_i u_i^\intercal] + n\rho \mathbf{I}\right)^{-1} \sum_{i=1}^n \mathbb{E}_{u_i}[u_i u_i^\intercal]\tilde{x}_i = \frac{1}{3}\left(\frac{1}{3} + \rho\right)^{-1} \frac{1}{n}\sum_{i=1}^n \tilde{x}_i. \quad (37)$$

Suppose $p = 10$ and $\rho = 1$. We compare the performance of DSGD (3) and the centralized implementation (30) for solving problem (36) with the same stepsize policy $\alpha_k = 20/(k+20), \forall k$, and the same initial solutions: $x_i(0) = \mathbf{0}$, $\forall i$, (DSGD) and $x(0) = \mathbf{0}$ (SGD). It can be seen from (37) and the definition of $\tilde{x}_i$ that $\sum_{i=1}^n \|x_i(0) - x_*\|^2 = \mathcal{O}(n)$. Moreover, $\nabla f_i(x_*) = 2\mathbb{E}_{u_i,v_i}[(u_i^\intercal x_* - v_i)u_i] + 2\rho x_* = 2\mathbb{E}_{u_i}[u_i u_i^\intercal](x_* - \tilde{x}_i) + 2\rho x_* = \frac{2}{3}(x_* - \tilde{x}_i) + 2\rho x_*$. Therefore, we have $\sum_{i=1}^n \|\nabla f_i(x_*)\|^2 = \mathcal{O}(n)$.

In Fig. 1, we provide an illustration example that compares the performance of DSGD and SGD, assuming $n = 25$. For DSGD, we consider two different network topologies: ring network topology as shown in Fig. 2(a) and square grid network topology as shown in Fig. 2(a). For both network topologies, we use Metropolis weights for constructing the maxing matrix $\mathbf{W}$ (see [55]). It can be seen that DSGD performs asymptotically as well as SGD, while the time it takes for DSGD to catch up with SGD depends on the network topology. For grid networks which are better connected than rings, the corresponding transient time is shorter.

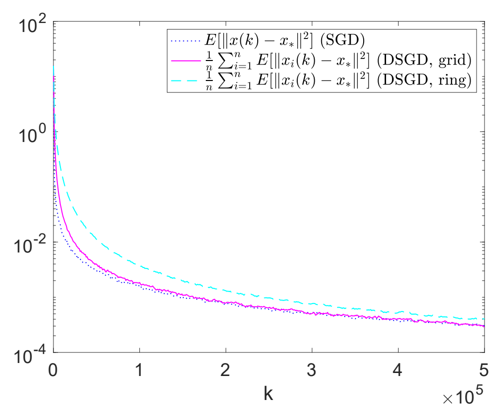
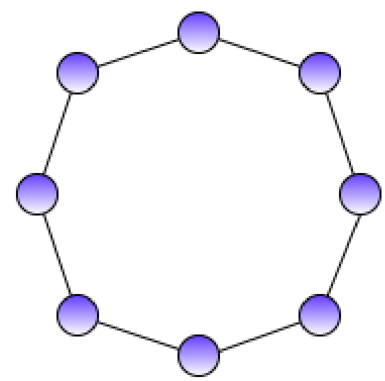
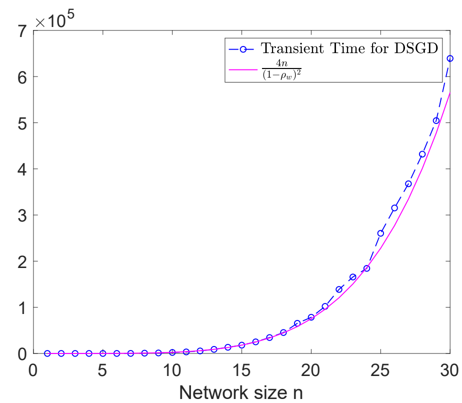


Fig. 1. The performance comparison between DSGD and SGD for online Ridge regression ($n = 25$). The results are averaged over 200 Monte Carlo simulation.

To further verify the conclusions of Corollary 2 and Theorem 3, we define the transient time for DSGD as $\inf\{k : \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|x_i(k) - x_*\|^2] \le 2\mathbb{E}[\|x(k) - x_*\|^2]\}$. For DSGD, we first assume a ring network topology and plot the transient times for DSGD and $\frac{4n}{(1-\rho_w)^2}$ as a function of the network size n in Fig. 2(b). We then consider a square grid network topology as shown in Fig. 2(a) and plot the transient times for DSGD and $\frac{7n}{(1-\rho_w)^2}$ in Fig. 2 (b). It can be seen that the two curves in Fig. 2(b) and Fig. 3(b) are close to each other, respectively. This verifies the sharpness of Corollary 2.



(a) Ring network topology.



(b) Transient times for the ring network topology.

Fig. 2. Comparison of the transient times for DSGD and $\frac{4n}{(1-\rho_w)^2}$ as a function of the network size n for the ring network topology. The expected errors are approximated by averaging over 200 simulation results.

B. Logistic Regression

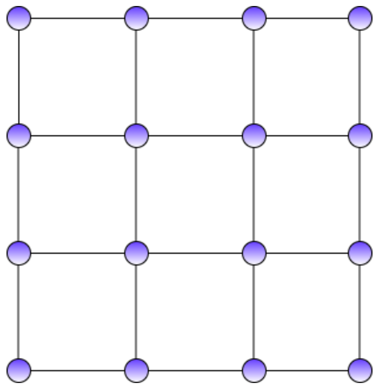
Consider the problem of classification on the MNIST dataset of handwritten digits (http://yann.lecun.com/exdb/mnist/). In particular, we classify digits 1 and 2 using logistic regression.[5] There are 12700 data points in total where each data point is a pair (u, v) with $u \in \mathbb{R}^{785}$ being the image input and $v \in \{0, 1\}$ being the label.[6]

Suppose each agent $i \in \mathcal{N}$ possesses a distinct local dataset $\mathcal{S}_i$ that is randomly taken from the database. To apply logistic regression for classification, we solve the following optimization problem based on all the agents' local datasets:

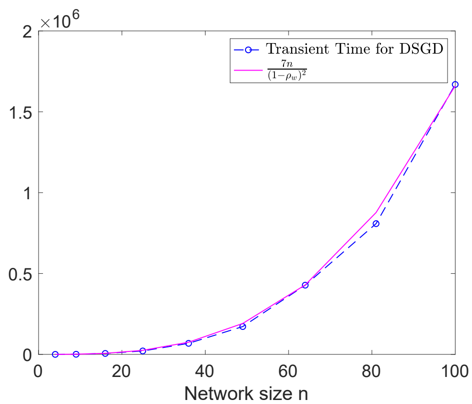
$$\min_{x \in \mathbb{R}^{785}} f(x) = \frac{1}{n}\sum_{i=1}^n f_i(x), \quad (38)$$

[5]The problem can be extended to classifying all 10 handwritten digits with multinomial logistic regression.

[6]Digit 1 is represented by label 0 and digit 2 is represented by label 1.



(a) Square grid network topology.



(b) Transient times for the square grid network topology.

Fig. 3. Comparison of the transient times for DSGD and $\frac{7n}{(1-\rho_w)^2}$ as a function of the network size n for the square grid network topology ($n = 4, 9, 16, 25, 36, 49, 64, 81, 100$). The expected errors are approximated by averaging over 200 simulation results.

where

$$f_i(x) := \frac{1}{|\mathcal{S}_i|} \sum_{j \in \mathcal{S}_i} \left[\log(1 + \exp(-x^\intercal u_j)) + (1 - v_j) x^\intercal u_j\right] + \frac{\lambda}{2}\|x\|^2,$$

where λ is the regularization parameter.[7] Given any solution x, agent i is able to compute an unbiased estimate of $\nabla f_i(x)$ using one (or a minibatch of) randomly chosen data point (u_i, v_i) from $\mathcal{S}_i$, that is,

$$g_i(x, u_i, v_i) = \frac{-u_j}{1 + \exp(x^\intercal u_j)} + (1 - v_j)u_j + \lambda x.$$

In the experiments, suppose each local dataset $\mathcal{S}_i$ contains 50 data points, and $\lambda = 1$. At each iteration of the DSGD algorithm, agent i computes a stochastic gradient of $f_i(x_i(k))$ with one randomly chosen data point from $\mathcal{S}_i$. We compare the performance of DSGD (3) and centralized SGD (30) for solving problem (38) with the same stepsize policy $\alpha_k = 6/(k + 20), \forall k$, and the same initial solutions: $x_i(0) = \mathbf{0}$, $\forall i$, (DSGD) and $x(0) = \mathbf{0}$ (SGD). It can be numerically verified that $\sum_{i=1}^n \|x_i(0) - x_*\|^2 = \mathcal{O}(n)$ and $\sum_{i=1}^n \|\nabla f_i(x_*)\|^2 = \mathcal{O}(n)$.

[7]The obtained optimal solution x_* of problem (38) can then be used for predicting the label for any image input u through the decision function $h(u) := \frac{1}{1+\exp(-x_*^\intercal u)}$.

The transient time for DSGD is defined in the same way as in the ridge regression example. In Fig. 4 and Fig. 5, we plot the transient times for DSGD as a function of the network size n for ring and grid networks, respectively. We find that the curves are close to $\frac{n}{4(1-\rho_w)^{1.5}}$, rather than a multiple of $\frac{n}{(1-\rho_w)^2}$, implying that the experimental results are better than the theoretically derived worst-case performance given in Corollary 2. Hence in practice, the performance of the DSGD algorithm depends on the specific problem instances and can be better than the worst-case situation in terms of transient times.

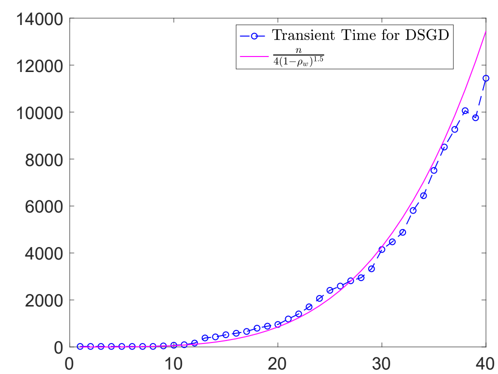


Fig. 4. Comparison of the transient times for DSGD and $\frac{n}{4(1-\rho_w)^{1.5}}$ as a function of the network size n for the ring network topology. The expected errors are approximated by averaging over 200 simulation results.

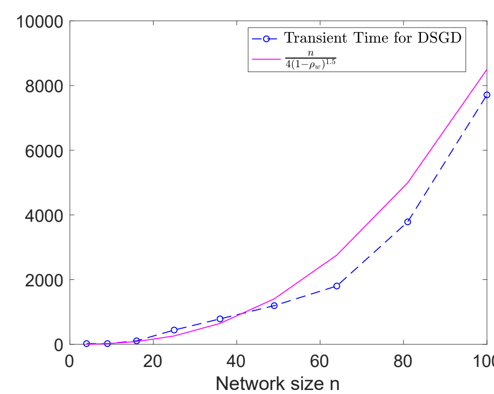


Fig. 5. Comparison of the transient times for DSGD and $\frac{n}{4(1-\rho_w)^{1.5}}$ as a function of the network size n for the grid network topology ($n = 4, 9, 16, 25, 36, 49, 64, 81, 100$). The expected errors are approximated by averaging over 200 simulation results.

VI. Conclusions

This paper is devoted to the non-asymptotic analysis of network independence for the distributed stochastic gradient descent (DSGD) method. We show that in expectation, the algorithm asymptotically achieves the optimal network independent convergence rate compared to SGD, and identify the non-asymptotic convergence rate as a function of characteristics of the objective functions and the network. In addition, we compute the time needed for DSGD to reach its asymptotic rate of convergence and prove the sharpness of the obtained result. Future work will consider more general problems such as nonconvex objectives and constrained optimization. It will also be of interest to explore the transient times of asynchronous distributed stochastic gradient algorithms which enjoy greater flexibility and less communication overhead.

REFERENCES

[1] J. Chen and A. H. Sayed, "Diffusion adaptation strategies for distributed optimization and learning over networks," *IEEE Transactions on Signal Processing*, vol. 60, no. 8, pp. 4289–4305, 2012.
[2] A. I. Forrester, A. Sóbester, and A. J. Keane, "Multi-fidelity optimization via surrogate modelling," *Proceedings of the Royal Society of London A*, vol. 463, no. 2088, pp. 3251–3269, 2007.
[3] A. Nedić, A. Olshevsky, and C. A. Uribe, "Fast convergence rates for distributed non-bayesian learning," *IEEE Transactions on Automatic Control*, vol. 62, no. 11, pp. 5538–5553, 2017.
[4] K. Cohen, A. Nedić, and R. Srikant, "On projected stochastic gradient descent algorithm with weighted averaging for least squares regression," *IEEE Transactions on Automatic Control*, vol. 62, no. 11, pp. 5974–5981, 2017.
[5] B. Baingana, G. Mateos, and G. B. Giannakis, "Proximal-gradient algorithms for tracking cascades over social networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 4, pp. 563–575, 2014.
[6] B. Ying, K. Yuan, and A. H. Sayed, "Supervised learning under distributed features," *IEEE Transactions on Signal Processing*, vol. 67, no. 4, pp. 977–992, 2018.
[7] S. A. Alghunaim and A. H. Sayed, "Distributed coupled multi-agent stochastic optimization," *IEEE Transactions on Automatic Control*, 2019.
[8] T. S. Brisimi, R. Chen, T. Mela, A. Olshevsky, I. C. Paschalidis, and W. Shi, "Federated learning of predictive models from federated electronic health records," *International journal of medical informatics*, vol. 112, pp. 59–67, 2018.
[9] K. Cohen, A. Nedić, and R. Srikant, "Distributed learning algorithms for spectrum sharing in spatial random access wireless networks," *IEEE Transactions on Automatic Control*, vol. 62, no. 6, pp. 2854–2869, 2017.
[10] G. Mateos and G. B. Giannakis, "Distributed recursive least-squares: Stability and performance analysis," *IEEE Transactions on Signal Processing*, vol. 60, no. 7, pp. 3740–3754, 2012.
[11] A. Reisizadeh, A. Mokhtari, H. Hassani, and R. Pedarsani, "An exact quantized decentralized gradient descent algorithm," *IEEE Transactions on Signal Processing*, vol. 67, no. 19, pp. 4934–4947, 2019.
[12] J. Tsitsiklis, D. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE transactions on automatic control*, vol. 31, no. 9, pp. 803–812, 1986.
[13] A. Nedić and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.
[14] A. Nedić, A. Ozdaglar, and P. A. Parrilo, "Constrained consensus and optimization in multi-agent networks," *IEEE Transactions on Automatic Control*, vol. 55, no. 4, pp. 922–938, 2010.
[15] I. Lobel, A. Ozdaglar, and D. Feijer, "Distributed multi-agent optimization with state-dependent communication," *Mathematical programming*, vol. 129, no. 2, pp. 255–284, 2011.
[16] 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.
[17] S. S. Kia, J. Cortés, and S. Martínez, "Distributed convex optimization via continuous-time coordination algorithms with discrete-time communication," *Automatica*, vol. 55, pp. 254–264, 2015.
[18] 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.
[19] P. Di Lorenzo and G. Scutari, "Next: In-network nonconvex optimization," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 2, no. 2, pp. 120–136, 2016.
[20] G. Qu and N. Li, "Harnessing smoothness to accelerate distributed optimization," *IEEE Transactions on Control of Network Systems*, 2017.
[21] A. Nedić, A. Olshevsky, and W. Shi, "Achieving geometric convergence for distributed optimization over time-varying graphs," *SIAM Journal on Optimization*, vol. 27, no. 4, pp. 2597–2633, 2017.
[22] J. Xu, S. Zhu, Y. C. Soh, and L. Xie, "Convergence of asynchronous distributed gradient methods over stochastic networks," *IEEE Transactions on Automatic Control*, vol. 63, no. 2, pp. 434–448, 2017.
[23] S. Pu, W. Shi, J. Xu, and A. Nedić, "Push-pull gradient methods for distributed optimization in networks," *IEEE Transactions on Automatic Control*, 2020.
[24] J. Chen and A. H. Sayed, "On the limiting behavior of distributed optimization strategies," in *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2012, pp. 1535–1542.
[25] ——, "On the learning behavior of adaptive networks—part i: Transient analysis," *IEEE Transactions on Information Theory*, vol. 61, no. 6, pp. 3487–3517, 2015.
[26] ——, "On the learning behavior of adaptive networks—part ii: Performance analysis," *IEEE Transactions on Information Theory*, vol. 61, no. 6, pp. 3518–3548, 2015.
[27] H. Robbins and S. Monro, "A stochastic approximation method," *The annals of mathematical statistics*, pp. 400–407, 1951.
[28] J. Kiefer, J. Wolfowitz *et al.*, "Stochastic estimation of the maximum of a regression function," *The Annals of Mathematical Statistics*, vol. 23, no. 3, pp. 462–466, 1952.
[29] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM Journal on optimization*, vol. 19, no. 4, pp. 1574–1609, 2009.
[30] 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.
[31] J. C. Duchi, A. Agarwal, and M. J. Wainwright, "Dual averaging for distributed optimization: Convergence analysis and network scaling," *IEEE Transactions on Automatic control*, vol. 57, no. 3, pp. 592–606, 2012.
[32] P. Bianchi and J. Jakubowicz, "Convergence of a multi-agent projected stochastic gradient algorithm for non-convex optimization," *IEEE Transactions on Automatic Control*, vol. 58, no. 2, pp. 391–405, 2013.
[33] Z. J. Towfic and A. H. Sayed, "Adaptive penalty-based distributed stochastic convex optimization," *Signal Processing, IEEE Transactions on*, vol. 62, no. 15, pp. 3924–3938, 2014.
[34] N. Chatzipanagiotis and M. M. Zavlanos, "A distributed algorithm for convex constrained optimization under noise," *IEEE Transactions on Automatic Control*, vol. 61, no. 9, pp. 2496–2511, 2016.
[35] A. Nedić and A. Olshevsky, "Stochastic gradient-push for strongly convex functions on time-varying directed graphs," *IEEE Transactions on Automatic Control*, vol. 61, no. 12, pp. 3936–3947, 2016.
[36] M. O. Sayin, N. D. Vanli, S. S. Kozat, and T. Başar, "Stochastic subgradient algorithms for strongly convex optimization over distributed networks," *IEEE Transactions on Network Science and Engineering*, vol. 4, no. 4, pp. 248–260, 2017.
[37] G. Lan, S. Lee, and Y. Zhou, "Communication-efficient algorithms for decentralized and stochastic optimization," *Mathematical Programming*, pp. 1–48, 2017.
[38] B. Sirb and X. Ye, "Decentralized consensus algorithm with delayed and stochastic gradients," *SIAM Journal on Optimization*, vol. 28, no. 2, pp. 1232–1254, 2018.
[39] D. Jakovetic, D. Bajovic, A. K. Sahu, and S. Kar, "Convergence rates for distributed stochastic optimization over random networks," in *2018 IEEE Conference on Decision and Control (CDC)*. IEEE, 2018, pp. 4238–4245.
[40] R. Xin, U. A. Khan, and S. Kar, "Variance-reduced decentralized stochastic optimization with gradient tracking," *arXiv preprint arXiv:1909.11774*, 2019.
[41] G. Morral, P. Bianchi, and G. Fort, "Success and failure of adaptation-diffusion algorithms for consensus in multi-agent networks," in *53rd IEEE Conference on Decision and Control*. IEEE, 2014, pp. 1476–1481.
[42] ——, "Success and failure of adaptation-diffusion algorithms with decaying step size in multiagent networks," *IEEE Transactions on Signal Processing*, vol. 65, no. 11, pp. 2798–2813, 2017.
[43] Z. J. Towfic, J. Chen, and A. H. Sayed, "Excess-risk of distributed stochastic learners," *IEEE Transactions on Information Theory*, vol. 62, no. 10, pp. 5753–5785, 2016.
[44] S. Pu and A. Garcia, "A flocking-based approach for distributed stochastic optimization," *Operations Research*, vol. 1, pp. 267–281, 2018.

[45] ——, "Swarming for faster convergence in stochastic optimization," *SIAM Journal on Control and Optimization*, vol. 56, no. 4, pp. 2997–3020, 2018.

[46] 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. 5336–5346.

[47] M. Assran, N. Loizou, N. Ballas, and M. Rabbat, "Stochastic gradient push for distributed deep learning," in *International Conference on Machine Learning*. PMLR, 2019, pp. 344–353.

[48] S. Pu and A. Nedić, "Distributed stochastic gradient tracking methods," *Mathematical Programming*, vol. 187, no. 1, pp. 409–457, 2021.

[49] A. Spiridonoff, A. Olshevsky, and I. C. Paschalidis, "Robust asynchronous stochastic gradient-push: Asymptotically optimal and network-independent performance for strongly convex functions." *Journal of Machine Learning Research*, vol. 21, no. 58, pp. 1–47, 2020.

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

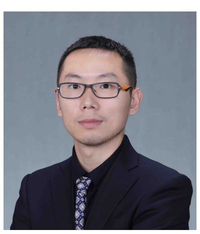
[51] S. Pu, A. Olshevsky, and I. C. Paschalidis, "Asymptotic network independence in distributed stochastic optimization for machine learning: Examining distributed and centralized stochastic gradient descent," *IEEE signal processing magazine*, vol. 37, no. 3, pp. 114–122, 2020.

[52] D. Williams, *Probability with martingales*. Cambridge university press, 1991.

[53] A. Rakhlin, O. Shamir, and K. Sridharan, "Making gradient descent optimal for strongly convex stochastic optimization," in *Proceedings of the 29th International Coference on International Conference on Machine Learning*. Omnipress, 2012, pp. 1571–1578.

[54] A. Olshevsky, "Linear time average consensus and distributed optimization on fixed graphs," *SIAM Journal on Control and Optimization*, vol. 55, no. 6, pp. 3990–4014, 2017.

[55] A. Nedić, A. Olshevsky, and M. G. Rabbat, "Network topology and communication-computation tradeoffs in decentralized optimization," *Proceedings of the IEEE*, vol. 106, no. 5, pp. 953–976, 2018.



Shi Pu is currently an assistant professor in the School of Data Science, The Chinese University of Hong Kong, Shenzhen, China. He is also affiliated with Shenzhen Research Institute of Big Data. He received a B.S. Degree from Peking University, in 2012, and a Ph.D. Degree in Systems Engineering from the University of Virginia, in 2016. He was a postdoctoral associate at the University of Florida, from 2016 to 2017, a postdoctoral scholar at Arizona State University, from 2017 to 2018, and a postdoctoral associate at Boston University, from 2018 to 2019. His research interests include distributed optimization, network science, machine learning, and game theory.

Alex Olshevsky received the B.S. degrees in applied mathematics and electrical engineering from Georgia Tech and the Ph.D. degree in EECS from MIT. He is currently an Associate Professor at the ECE department at Boston University. Dr. Olshevsky is a recipient of the NSF CAREER Award, the AFOSR Young Investigator Award, the INFORMS Prize for the best paper on the interface of operations research and computer science, a SIAM Award for annual paper from the SIAM Journal on Control and Optimization chosen to be reprinted in SIAM Review, and an IMIA award for best paper on clinical informatics.

Ioannis Ch. Paschalidis (M'96–SM'06–F'14) received a Ph.D. in EECS from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 1996. He is a Professor at Boston University, Boston, MA and the Director of the Center for Information and Systems Engineering. His research interests lie in the fields of systems and control, optimization, stochastic systems, machine learning, and computational biology and medicine. He is a recipient of the NSF CAREER award, several best paper and best algorithmic performance awards, and a 2014 IBM/IEEE Smarter Planet Challenge Award. He was an invited participant at the 2002 Frontiers of Engineering Symposium, organized by the U.S. National Academy of Engineering and the 2014 U.S. National Academies Keck Futures Initiative (NAFKI) Conference. During 2013–2019 he was the founding Editor-in-Chief of the IEEE Transactions on Control of Network Systems.