SIAM J. OPTIM.
Vol. 33, No. 2, pp. 978–1005
© 2023 Society for Industrial and Applied Mathematics

Downloaded 06/30/23 to 128.101.167.180 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

DIMIX: DIMINISHING MIXING FOR SLOPPY AGENTS*

HADI REISIZADEH†, BEHROUZ TOURI‡, AND SOHEIL MOHAJER†

Abstract. We study nonconvex distributed optimization problems where a set of agents collaboratively solve a separable optimization problem that is distributed over a time-varying network. The existing methods to solve these problems rely on (at most) one-time-scale algorithms, where each agent performs a diminishing or constant step-size gradient descent at the average estimate of the agents in the network. However, if possible at all, exchanging exact information, which is required to evaluate these average estimates, potentially introduces a massive communication overhead. Therefore, a reasonable practical assumption to be made is that agents only receive a rough approximation of the neighboring agents' information. To address this, we introduce and study a *two-time-scale* decentralized algorithm with a broad class of *lossy* information sharing methods (which includes noisy, quantized, and/or compressed information sharing) over *time-varying* networks. In our method, one time-scale suppresses the (imperfect) incoming information from the neighboring agents, and one time-scale operates on local cost functions' gradients. We show that with a proper choices for the step-sizes' parameters, the algorithm achieves a convergence rate of $\mathcal{O}(T^{-1/3+\epsilon})$ for nonconvex distributed optimization problems over time-varying networks for any $\epsilon > 0$.

Key words. nonconvex optimization, time-varying graphs, distributed multiagent system, distributed optimization, gradient descent algorithms

MSC code. 90Cxx

DOI. 10.1137/21M143546X

1. Introduction.

Distributed learning serves as a learning framework where a set of computing nodes/agents are interested in collaboratively solving an optimization problem. In this paradigm, in the absence of a central node, the learning task solely depends on on-device computation and local communication among the neighboring agents. With the appearance of modern computation architectures and the decentralized nature of storage, large-scale distributed computation frameworks have received significant attention due to data locality, privacy, data ownership, and scalability to larger datasets and systems. These features of distributed learning have led to applications in several domains including distributed deep networks [8, 1, 15], distributed sensor networks [28, 16], and network resource allocation [31, 7].

Related works. Decentralized consensus or averaging-based optimization algorithms have been studied extensively over the past few years [25, 22, 6, 39, 14]. It has been shown that when a fixed step-size is utilized, the loss function decreases with the rate of $\mathcal{O}(1/T)$ until the estimates reach a neighborhood of the (local) minimum of the objective cost function [25]. However, with a fixed step-size, the local estimates may not converge to an optimal point [39]. To remedy this, the diminishing step-size

* Received by the editors July 23, 2021; accepted for publication (in revised form) November 20, 2022; published electronically June 22, 2023.
https://doi.org/10.1137/21M143546X
Funding: The work of the first and third authors is supported in part by the National Science Foundation (NSF) under grant CCF-1749981.

† Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55414 USA (hadir@umn.edu, soheil@umn.edu).

‡ Department of Electrical and Computer Engineering, University of California San Diego, La Jolla, CA 92093 USA (btouri@ucsd.edu).

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 06/30/23 to 128.101.167.180 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

variation of the algorithm is introduced and studied for convex [26, 14] and nonconvex problems [32, 34, 40].

A majority of the existing works in this area suffer from requiring large communication overhead, as a (potentially) massive amount of local information is needed to be exchanged among the agents throughout the learning algorithm. In addition, such a communication load can lead to a major delay in the convergence of the algorithm and become a severe bottleneck when the dimension of the model is large. To mitigate the communication overhead, several compression techniques are introduced to reduce the size of the exchanged information. In particular, these techniques have been used in the context of the distributed averaging/consensus problem [17, 25, 5], where each node has an initial numerical value and aims at evaluating the average of all initial values using exchange of quantized information, over a fixed or a time-varying network.

For distributed optimization, an inexact proximal gradient method with a fixed step-size over a fixed network for strongly convex loss functions is proposed in [27]. In this algorithm, each node applies an adaptive deterministic quantizer on its model and gradient before sending them to its neighbors. It is shown that under certain conditions, the algorithm provides a linear convergence rate. In a related work [29], the authors proposed a decentralized gradient descent method, named QuanTimed-DSGD, with fixed step-sizes. In this algorithm, agents exchange a quantized version of their models/estimates in order to reduce the communication load (see Remark 2.6 for more details). A time-varying version of this algorithm with vanishing step-sizes is studied in [30] for strongly convex objective functions. To compensate the quantization error, a decentralized (diminishing) gradient descent algorithm is proposed in [19, 18] using error-feedback. The proposed algorithm achieves the convergence rate of $\mathcal{O}(T^{-1})$ and $\mathcal{O}(T^{-1/2})$ for strongly and nonconvex objective functions, respectively. However, the nature of the algorithm restricts its use to time-invariant networks, and in addition, the feedback mechanism cannot compensate communication noise between the nodes. In [33], a two-time-scale gradient descent algorithm was presented for distributed constrained and convex optimization problems over an independent and identically distributed (i.i.d.) communication graph with noisy communication links, and subgradient errors. It is shown that if the random communication satisfies certain conditions, proper choices of the time-scale parameters result in the almost sure convergence of the local states to an optimal point. Another interesting approach to address exact convergence for distributed optimization with fixed gradient step-sizes under a *noiseless* communication model is to use gradient tracking methods [10, 38].

Our contributions. We introduce and study a *two-time-scale* decentralized gradient descent algorithm for a *broad* class of imperfect sharing of information over *time-varying communication networks* for distributed optimization problems with smooth *nonconvex* local cost functions. Here, one time-scale addresses the convergence of the local estimates to a stationary point while the second time-scale is introduced to suppress the noisy effect of the imperfect incoming information from the neighbors.

Our main result shows that with a proper choice of the parameters for the two diminishing step-size sequences, the temporal average of the expected norm of the gradients decreases with the rate of $\mathcal{O}(T^{-1/3+\epsilon})$. To prove this result, we have established new techniques to analyze the interplay between the two time-scales, in particular in the presence of underlying time-varying networks.

Paper organization. After introducing some notation, we present the problem formulation, the algorithm, the main result, and some of its practical implications in section 2. To prove the main result, we first provide some auxiliary lemmas in

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 06/30/23 to 128.101.167.180 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

section 3, whose proofs are presented in Appendix B. In section 4, we present the proof of the main result. Finally, we conclude the paper in section 5.

Notation and basic terminology. Throughout this paper, we use $[n]$ to denote the set of integers $\{1, 2, \dots, n\}$. In this paper, we are dealing with n agents that are minimizing a function in $\mathbb{R}^d$. For notational convenience, throughout this paper, we assume that the underlying functions are acting on *row* vectors, and hence, we view vectors in $\mathbb{R}^{1\times d} = \mathbb{R}^d$ as row vectors. The rest of the vectors, i.e., the vectors in $\mathbb{R}^{n\times 1} = \mathbb{R}^n$, are assumed to be column vectors. The ℓ_2-norm of a vector $\mathbf{x} \in \mathbb{R}^d$ is defined as $\|\mathbf{x}\|^2 = \sum_{j=1}^d |x_j|^2$. The Frobenius norm of a matrix $A \in \mathbb{R}^{n\times d}$ is defined as $\|A\|_F^2 = \sum_{i=1}^n \|A_i\|^2 = \sum_{i=1}^n \sum_{j=1}^d |A_{ij}|^2$. A vector $\mathbf{r} \in \mathbb{R}^n$ is called stochastic if $r_i \geq 0$ and $\sum_{i=1}^n r_i = 1$. Similarly, a nonnegative matrix $A \in \mathbb{R}^{n\times d}$ is called (row) stochastic if $\sum_{j=1}^d A_{ij} = 1$ for every $i \in [n]$. For a matrix $A \in \mathbb{R}^{n\times d}$, we denote its ith row and jth column by A_i and A^j, respectively. Note that we deal with two types of vector throughout the paper. For an $n \times d$ matrix A and a strictly positive stochastic vector $\mathbf{r} \in \mathbb{R}^n$, we define the $\mathbf{r}$-norm of A by $\|A\|_{\mathbf{r}}^2 = \sum_{i=1}^n r_i \|A_i\|^2$. It can be verified that $\|\cdot\|_{\mathbf{r}}$ is a norm on the space of $n \times d$ matrices. Finally, we write $A \geq B$ if $A - B$ (is well-defined and) has only nonnegative entries.

2. Problem setup and main result.

In this section, first we formulate nonconvex distributed optimization problems over time-varying networks and introduce some standard assumptions on the underlying problem. After proposing our algorithm, we state our main result. Finally, we discuss the implications of our result on various important practical settings with imperfect information sharing.

2.1. Problem setup.

This paper is motivated by stochastic learning problems in which the goal is to solve

$$\min_{\mathbf{x}} L(\mathbf{x}) := \min_{\mathbf{x}} \mathbb{E}_{\xi\sim\mathcal{P}}[\ell(\mathbf{x}, \xi)], \tag{2.1}$$

where $\ell : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$ is a loss function, $\mathbf{x} \in \mathbb{R}^{1\times d} = \mathbb{R}^d$ is the decision/optimization *row* vector, and ξ is a random vector taking values in $\mathbb{R}^p$ that is drawn from an unknown underlying distribution $\mathcal{P}$. One of the key practical considerations that renders (2.1) as a challenging task is that the underlying distribution $\mathcal{P}$ is often unknown. Instead, we have access to N independent realizations of ξ and focus on solving the corresponding empirical risk minimization (ERM) problem, which is given by

$$\min_{\mathbf{x}} f(\mathbf{x}) := \min_{\mathbf{x}} \frac{1}{N} \sum_{j=1}^N \ell(\mathbf{x}, \xi_j), \tag{2.2}$$

where $f(\mathbf{x})$ is the empirical risk with respect to the data points $\mathcal{D} = \{\xi_1, \dots, \xi_N\}$. We assume that $\ell(\cdot, \cdot)$ is a nonconvex loss function, which potentially results in a nonconvex function $f(\cdot)$.

In distributed optimization, we have a network consisting of n computing nodes (agents, or workers), where each node i observes a nonoverlapping subset of $m_i = r_i N$ data points, denoted by $\mathcal{D}_i = \{\xi_1^i, \dots, \xi_{m_i}^i\}$, where $\mathcal{D} = \mathcal{D}_1 \cup \cdots \cup \mathcal{D}_n$. Here, r_i represents the fraction of the data that is processed at node $i \in [n]$. Note that the vector $\mathbf{r} = (r_1, \dots, r_n)$ is a strictly positive stochastic vector, i.e., $r_i > 0$ and $\sum_{i=1}^n r_i = 1$. Thus, the ERM problem in (2.2) can be written as the minimization of the weighted average of local empirical risk functions f_i for all nodes $i \in [n]$ in the network, i.e.,

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

$$\min_{\mathbf{x}} f(\mathbf{x}) = \min_{\mathbf{x}} \sum_{i=1}^{n} r_i f_i(\mathbf{x}) = \min_{\mathbf{x}} \frac{1}{N} \sum_{i=1}^{n} \sum_{\xi \in \mathcal{D}_i} \ell(\mathbf{x}, \xi), \tag{2.3}$$

where $f_i(\mathbf{x}) := \frac{1}{m_i} \sum_{\xi \in \mathcal{D}_i} \ell(\mathbf{x}, \xi) = \frac{1}{m_i} \sum_{j=1}^{m_i} \ell(\mathbf{x}, \xi_j^i)$. We can rewrite the ERM problem in (2.3) as a distributed consensus optimization problem, given by

$$\min_{\mathbf{x}_1, \ldots, \mathbf{x}_n} \sum_{i=1}^{n} r_i f_i(\mathbf{x}_i) \quad \text{subject to} \quad \mathbf{x}_1 = \mathbf{x}_2 = \cdots = \mathbf{x}_n. \tag{2.4}$$

Consider an $n \geq 2$ agents that are connected through a *time-varying* network. We represent this network at time $t \geq 1$ by the directed graph $\mathcal{G}(t) = ([n], \mathcal{E}(t))$, where the vertex set $[n]$ represents the set of agents and the edge set $\mathcal{E}(t) \subseteq \{(i,j) : i,j \in [n]\}$ represents the set of links at time t. At each discrete time $t \geq 1$, agent i can only send information to its (out-) neighbors in $\mathcal{E}(t)$, i.e., all $j \in [n]$ with $(i,j) \in \mathcal{E}(t)$.

To discuss our algorithm (`DIMIX`) for solving (2.4) distributively, let us first discuss its general structure and the required information at each node for its execution. In this algorithm, at each iteration $t \geq 1$, agent $i \in [n]$ updates its estimate $\mathbf{x}_i(t) \in \mathbb{R}^d$ of an optimizer of (2.3). To this end, it utilizes the gradient information of its own local cost function $f_i(\mathbf{x})$ as well as a *noisy/lossy* average of its current neighbors estimates, denoted by $\hat{\mathbf{x}}_i(t) := \sum_{j=1}^{n} W_{ij}(t)\mathbf{x}_j(t) + \mathbf{e}_i(t)$. Here, $W(t)$ is a *row-stochastic* matrix that is consistent with the underlying connectivity network $\mathcal{G}(t)$ (i.e., $W_{ij}(t) > 0$ only if $(j,i) \in \mathcal{E}(t)$) and $\mathbf{e}_i(t) \in \mathbb{R}^d$ is a random noise vector. Later, in section 2.4 we discuss several noisy and lossy information sharing architectures (quantization and noisy communication) that fit in this broad information structure.

Now we are ready to discuss the `DIMIX` algorithm. In this algorithm, using the information available to agent i at time t, agent i updates its current estimate by computing a *diminishing* weighted average of its own state and the noisy average of its neighbors' estimates, and moves along its local gradient. More formally, the update rule at node $i \in [n]$ is given by

$$\mathbf{x}_i(t+1) = (1 - \beta(t))\mathbf{x}_i(t) + \beta(t)\hat{\mathbf{x}}_i(t) - \alpha(t)\beta(t)\nabla f_i(\mathbf{x}_i(t)), \tag{2.5}$$

where $\alpha(t) = \frac{\alpha_0}{(t+\tau)^\nu}$ and $\beta(t) = \frac{\beta_0}{(t+\tau)^\mu}$ for some $\mu, \nu \in (0,1)$ are the diminishing step-sizes of the algorithm, and $\tau \geq 0$ is an arbitrary shift, that is introduced to accelerate the finite-time performance of the algorithm. For notational simplicity, let

$$X(t) := \begin{bmatrix} \mathbf{x}_1(t) \\ \vdots \\ \mathbf{x}_n(t) \end{bmatrix}, \quad E(t) := \begin{bmatrix} \mathbf{e}_1(t) \\ \vdots \\ \mathbf{e}_n(t) \end{bmatrix}, \quad \nabla f(X(t)) := \begin{bmatrix} \nabla f_1(\mathbf{x}_1(t)) \\ \vdots \\ \nabla f_n(\mathbf{x}_n(t)) \end{bmatrix}. \tag{2.6}$$

Since $\hat{\mathbf{x}}_i(x) = \sum_{j=1}^{n} W_{ij}(t)\mathbf{x}_j(t) + \mathbf{e}_i(t)$, we can rewrite the update rule in (2.5) in the matrix format

$$X(t+1) = ((1 - \beta(t))I + \beta(t)W(t))X(t) + \beta(t)E(t) - \alpha(t)\beta(t)\nabla f(X(t)). \tag{2.7}$$

Let us discuss some important aspects of the above update rule. Note that both $\alpha(t)$ and $\beta(t)$ are diminishing step-sizes. If $\beta(t) = \beta_0 < 1$ and $\alpha(t) = \alpha_0 < 1$ are both constants, then the dynamics in (2.7) reduces to the algorithm proposed in [29] for both convex and nonconvex cost functions. Alternatively, if $\beta(t) = \beta_0 < 1$ is a constant sequence and $E(t) = \mathbf{0}$ for all $t \geq 1$, (2.7) would be reduced to the averaging-based

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 06/30/23 to 128.101.167.180 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 06/30/23 to 128.101.167.180 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

distributed optimization with diminishing steps-sizes (for gradients), which is introduced and studied in [26] for local convex cost functions $f_i(\mathbf{x})$. The newly introduced time-scale/step-size $\beta(t)$ suppresses the incoming noise $\mathbf{e}_i(t)$ from the neighboring agents. However, $\beta(t)$ also suppresses the incoming signal level $\sum_{j=1}^{n} W_{ij}(t)\mathbf{x}_j(t)$ at each node i. This casts a major technical challenge for establishing convergence-to-consensus guarantees for the algorithm over time-varying networks. On the other hand, the diminishing step-size for the gradient update is $\hat{\alpha}(t) = \alpha(t)\beta(t)$. We chose to represent our algorithm in this way to ensure that the local mixing (consensus) scheme is operated on a faster time-scale than the gradient descent.

2.2. Assumptions. In order to provide performance guarantees for `DIMIX`, we need to assume certain regularity conditions on (i) the statistics of the (neighbors' averaging) noise process $\{E(t)\}$, (ii) the mixing properties of the weight sequence $\{W(t)\}$, and (iii) the loss function $\ell(\cdot,\cdot)$.

First, we discuss our main assumption on the noise sequence $\{E(t)\}$.

Assumption 1 (neighbors state estimate assumption). We suppose that $\{X(t)\}$ is adapted to a filtration $\{\mathcal{F}_t\}$ on the underlying probability space (see, e.g., section 5.2 in [12]). We assume that there exists some $\gamma > 0$ such that for all $i \in [n]$ and all $t \geq 1$, the noise sequence $\{\mathbf{e}_i(t)\}$ satisfies

$$\mathbb{E}[\mathbf{e}_i(t) \,|\, \mathcal{F}_t] = 0 \quad \text{and} \quad \mathbb{E}[\|\mathbf{e}_i(t)\|^2 \,|\, \mathcal{F}_t] \leq \gamma. \tag{2.8}$$

Note that the natural filtration of the random process $\{X(t)\}$ is one choice for $\{\mathcal{F}_t\}$. Thus, (2.8) reduces to $\mathbb{E}[\mathbf{e}_i(t) \,|\, X(1), \ldots, X(t)] = 0$ and $\mathbb{E}[\|\mathbf{e}_i(t)\|^2 \,|\, X(1), \ldots, X(t)] \leq \gamma$.

Next, we discuss the main assumption on the network connectivity which relates to information mixing over the time-varying network.

Assumption 2 (connectivity assumption). We assume that the weight matrix sequence $\{W(t)\}$ in (2.7) satisfies the following properties.

(a) Stochastic with common stationary distribution: For all $t \geq 1$, $W(t)$ is nonnegative, $W(t)\mathbf{1} = \mathbf{1}$, and $\mathbf{r}^T W(t) = \mathbf{r}^T$, where $\mathbf{1} \in \mathbb{R}^n$ is the all-one vector, and $\mathbf{r} > 0$ is the weight vector.
(b) Bounded nonzero elements: There exists some $\eta > 0$ such that $W_{ij}(t) > 0$ implies $W_{ij}(t) \geq \eta$ for all $i, j \in [n]$ and $t \geq 1$.
(c) B-connected: For a fixed integer $B \geq 1$, the graph $([n], \bigcup_{k=t+1}^{t+B} \mathcal{E}(k))$ is strongly connected for all $t \geq 1$, where $\mathcal{E}(k) = \{(j,i) \,|\, W_{ij}(k) > 0\}$.

Finally for the loss function $\ell(\cdot,\cdot)$, we make the following assumption.

Assumption 3 (stochastic loss function assumption). We assume that

(a) the function $\ell(\cdot,\cdot)$ is K-smooth with respect to its first argument, i.e., for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ and $\xi \in \mathcal{D}$ we have that $\|\nabla\ell(\mathbf{x},\xi) - \nabla\ell(\mathbf{y},\xi)\| \leq K\|\mathbf{x} - \mathbf{y}\|$;
(b) stochastic gradient $\nabla\ell(\mathbf{x},\xi)$ is unbiased and variance bounded, i.e.,

$$\mathbb{E}_\xi[\nabla\ell(\mathbf{x},\xi)] = \nabla L(\mathbf{x}), \quad \mathbb{E}_\xi[\|\nabla\ell(\mathbf{x},\xi) - \nabla L(\mathbf{x})\|^2] \leq \sigma^2.$$

Note that Assumption 3(b) implies a homogeneous sampling, i.e., each agent draws i.i.d. sample points from a data batch. In a related work [21], a stronger assumption has been considered which allows for heterogeneous data samples.

2.3. Main result. Here, we characterize the convergence rates of our algorithm for the K-smooth nonconvex loss functions. More precisely, we establish a rate for

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 06/30/23 to 128.101.167.180 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

the temporal average of the expected norm of the gradients for various choices of the time-scale parameters ν, μ.

THEOREM 2.1. *Suppose that Assumptions* 1–3 *hold and let* $\alpha(t) = \frac{\alpha_0}{(t+\tau)^\nu}$ *and* $\beta(t) = \frac{\beta_0}{(t+\tau)^\mu}$, *where* $\alpha_0, \beta_0 \in (0,1)$, $\tau \geq 0$, *and* $\nu, \mu \in (0,1)$ *are arbitrary constants with* $\mu \neq 1/2$ *and* $3\nu + \mu \neq 1$. *Then the weighted average estimates* $\bar{\mathbf{x}}(t) := \sum_{i=1}^n r_i \mathbf{x}_i(t)$ *generated by* (2.5) *satisfy*

$$M_\theta(\nu,\mu) := \left[\frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2]\right)^\theta\right]^{1/\theta} = \mathcal{O}\left(T^{-\min\{1-\nu-\mu,\mu-\nu,2\nu\}}\right), \tag{2.9}$$

where $\theta \in (0,1)$ *is an arbitrary constant.*

Furthermore, for $(\nu^\star, \mu^\star) = (\frac{1}{6}, \frac{1}{2})$ *we get the optimal rate of*

$$M_\theta(\nu^\star, \mu^\star) = \mathcal{O}\left(T^{-1/3}\ln T\right). \tag{2.10}$$

Remark 2.2. Note that the expectation operator $\mathbb{E}[\cdot]$ is over the randomness of the dataset $\mathcal{D}$ and the compression/communication noise. Moreover, note that the theorem above shows that the gradient of $f(\cdot)$ (which depends of the choice of $\mathcal{D}$) at the average state of $\overline{\mathbf{x}}(t)$ (which also depends on $\mathcal{D}$) vanishes at a certain rate. It is worth mentioning that this is not the performance of the average trajectory for the average function.

Remark 2.3. From (2.9), one has to maximize $\min\{1-\nu-\mu, \mu-\nu, 2\nu\}$ over $\nu, \mu \in (0,1)$ to achieve the fastest convergence for M_θ. This leads to $(\nu^\star, \mu^\star) = (1/6, 1/2)$, which none of the conditions $\mu \neq 1/2$ and $3\nu + \mu \neq 1$ hold for. However, one can choose $(\nu, \mu) = (1/6 + \epsilon/2, 1/2 + \epsilon/2)$ and obtain $M_\theta = \mathcal{O}(T^{-1/3+\epsilon})$ for any $\epsilon > 0$. Nevertheless, note that (2.10) provides a faster convergence rate of $\mathcal{O}(T^{-1/3}\ln T)$ for $(\nu^\star, \mu^\star) = (1/6, 1/3)$.

PROPOSITION 2.4. *Under the conditions of Theorem* 2.1, *for the optimum choice of* $(\nu^\star, \mu^\star) = (1/6, 1/3)$, *we have*

$$M_1(\nu^\star, \mu^\star) := \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2] \leq \mathcal{O}\left(T^{-1/3+\epsilon}\right) \tag{2.11}$$

for any $\epsilon > 0$. *Furthermore, in this case, for each agent* $i \in [n]$ *the convergence rate to consensus is given by*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{x}_i(t) - \bar{\mathbf{x}}(t)\|^2] \leq \mathcal{O}\left(T^{-1/3+\epsilon}\right). \tag{2.12}$$

As a result, combining (2.11), (2.12), *and Assumption* 3, *for all* $i \in [n]$, *we have*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\mathbf{x}_i(t))\|^2] \leq \mathcal{O}\left(T^{-1/3+\epsilon}\right).$$

Remark 2.5. We should comment that almost all the existing results and algorithms on distributed optimization algorithms for time-varying graphs assume a uniform positive lower bound on nonzero elements of the (effective) weight matrices [9, 25, 23, 36, 35, 39, 34, 2]. Absence of such an assumption significantly increases the

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

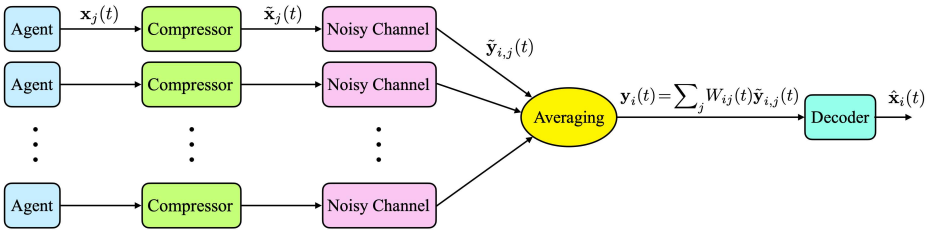

FIG. 1. *A general architecture for lossy information model.*

complexity of the convergence analysis of the algorithm. In our work, even though the stochastic matrix sequence $\{W(t)\}$ is assumed to be B-connected, the *effective averaging* weight sequence, given by $\{(1-\beta(t))I+\beta(t)W(t)\}$, has vanishing weights. One of the major theoretical contributions of this work is to introduce tools and techniques to study distributed optimization with diminishing weight sequences.

Remark 2.6. In a related work [29] on distributed optimization with compressed information sharing among the nodes, authors considered a fixed step-size (zero time-scale) version of our dynamics (2.5) with a fixed averaging matrix W. It is shown that for a given *termination time* T, the algorithm's step-sizes can be chosen (depending on T) such that the temporal average (up to iteration T) of the expected norm of the gradient (i.e., M_1 defined in (2.11)) does not exceed $c(T^{-1/3})$ (where $c>0$ is a constant). However, the algorithm needs to be re-executed with re-evaluated step-sizes if one targets another termination time T'. In this work, we use vanishing step-sizes $\alpha(t)$ and $\beta(t)$ (which do not depend on the termination time) and show that the same temporal average vanishes at the rate of $\mathcal{O}(T^{-1/3+\epsilon})$ for *every* iteration T and any arbitrarily small $\epsilon>0$.

2.4. Examples for stochastic noisy state estimation. The noisy information in (2.5) is very general and captures several models of imperfect data used in practice and/or theoretical studies. A rather general architecture that leads to such noisy/lossy estimates is demonstrated in Figure 1: Once the estimate $\mathbf{x}_j(t)$ of node j at iteration t is evaluated, node j may apply an operation (such as compression/sparsification or quantization) on its own model to generate $\tilde{\mathbf{x}}_j(t)$. This vector is sent over a potentially noisy communication channel, and a neighbor node i receives a corrupted version of $\tilde{\mathbf{x}}_j(t)$, say, $\tilde{\mathbf{y}}_{i,j}(t)$ from every neighbor node j. Upon collecting all channel outputs from its neighbors, node i computes their weighted average $\mathbf{y}_i(t)=\sum_{j=1}^n W_{ij}(t)\tilde{\mathbf{y}}_{i,j}(t)$ and decodes it to the approximate average model $\hat{\mathbf{x}}_i(t)$. In the following we describe three popular frameworks, in which each node i can only use an imperfect neighbors' average $\hat{\mathbf{x}}_i(t)$ to update its estimate. It is worth emphasizing that these are just some examples that lie under the general model in (2.5).

Example 1 (stochastic quantizer with bounded trajectory). The stochastic quantizer with a number of quantization levels s maps a vector $\mathbf{x}\in\mathbb{R}^d$ to a random vector $Q_s^S(\mathbf{x})\in\mathbb{R}^d$, where its ℓth entry is given by

$$\left[Q_s^S(\mathbf{x})\right]_\ell := \|\mathbf{x}\|\cdot\operatorname{sgn}(x_\ell)\cdot\zeta\left(|x_\ell|/\|\mathbf{x}\|,s\right),\quad \ell\in[d], \tag{2.13}$$

and $\zeta(x,s)$ is a random variable taking values

$$\zeta(x,s)=\begin{cases}\lceil sx\rceil/s & \text{w.p. } sx-\lfloor sx\rfloor,\\ \lfloor sx\rfloor/s & \text{w.p. } \lceil sx\rceil-sx.\end{cases} \tag{2.14}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 06/30/23 to 128.101.167.180 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 06/30/23 to 128.101.167.180 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Note that random variables $\{\zeta(\cdot,\cdot)\}$ are independent across the coordinates, agents, and time steps. Thus, in this case, the relationship between $\tilde{\mathbf{x}}_j(t)$ and $\mathbf{x}_j(t)$ in Figure 1 would be $\tilde{\mathbf{x}}_j(t) = Q_s^S(\mathbf{x}_j(t))$. Furthermore, the noisy channel is perfect, and the decoder component is just an identity function, i.e., $\mathbf{y}_{i,j}(t) = \tilde{\mathbf{x}}_j(t)$ and $\hat{\mathbf{x}}_i(t) = \mathbf{y}_i(t)$. It is shown in [3] that the output of this quantizer for an input $\mathbf{x} \in \mathbb{R}^d$ with a bounded norm $\|\mathbf{x}\|^2 \le D$ satisfies $\mathbb{E}[Q_s^S(\mathbf{x})] = \mathbf{x}$ and $\mathbb{E}[\|Q_s^S(\mathbf{x}) - \mathbf{x}\|^2] \le \min(\frac{\sqrt{d}}{s}, \frac{d}{s^2})D$. Therefore, the neighbors estimate for node i will be

$$\hat{\mathbf{x}}_i(t) = \sum_{j=1}^n W_{ij}(t) Q_s^S(\mathbf{x}_j(t)) = \sum_{j=1}^n W_{ij}(t)\mathbf{x}_j(t) + \mathbf{e}_i(t),$$

where $\mathbf{e}_i(t) = \sum_{j=1}^n W_{ij}(t)(Q_b^S(\mathbf{x}_j(t)) - \mathbf{x}_j(t))$ satisfies $\mathbb{E}[\mathbf{e}_i(t)|\mathcal{F}_t] = 0$ and

$$\mathbb{E}[\|\mathbf{e}_i(t)\|^2|\mathcal{F}_t] = \min\left(\sqrt{d}/s, d/s^2\right) D \sum_{j=1}^n W_{ij}^2(t) \le \min\left(\sqrt{d}/s, d/s^2\right) D,$$

provided that $\|\mathbf{x}_j(t)\|^2 \le D$ for every $j \in [n]$ and every $t \ge 1$. Therefore, the conditions of Assumption 1 are satisfied. Note that $\mathbf{e}_i(t)$ and $\mathbf{e}_j(t)$ might be correlated, especially when nodes i and j have common neighbor(s). However, this does not violate the conditions of Assumption 3. The bounded variance noise sequence, which is ensured by bounded trajectory here, has been implicitly and explicitly assumed in many related works [29, 37, 20, 11]. In addition, the above stochastic quantizer implicitly assumes a bounded trajectory as, otherwise, the state norm ($\|\mathbf{x}\|$, whose communication cost is ignored) requires infinite bits to be transmitted.

Example 2 (noisy communication). The noisy neighbor estimate model may arise due to imperfect communication between the agents. Consider a wireless network, in which the computing nodes communicate with their neighbors over a Gaussian channel, i.e., when node j sends its state $\mathbf{x}_j(t)$ (without compression, i.e., $\tilde{\mathbf{x}}_j(t) = \mathbf{x}_j(t)$) to its neighbor i, the signal received at node i is $\tilde{\mathbf{y}}_{i,j}(t) = \mathbf{x}_j(t) + \mathbf{z}_{i,j}(t)$, where $\mathbf{z}_{i,j}(t)$ is a zero-mean Gaussian noise with variance ζ^2, independent across (i,j), and t. Applying an identity map decoder at node i (i.e., $\hat{\mathbf{x}}_i(t) = \mathbf{y}_i(t)$) we have

$$\hat{\mathbf{x}}_i(t) = \sum_{j=1}^n W_{ij}(t)\left(\mathbf{x}_j(t) + \mathbf{z}_{i,j}(t)\right) = \sum_{j=1}^n W_{ij}(t)\mathbf{x}_j(t) + \sum_{j=1}^n W_{ij}(t)\mathbf{z}_{i,j}(t).$$

Therefore, we have $\mathbf{e}_i(t) = \sum_{j=1}^n W_{ij}(t)\mathbf{z}_{i,j}(t)$, from which we conclude $\mathbb{E}[\mathbf{e}_i(t)|\mathcal{F}_t] = 0$ and $\mathbb{E}[\|\mathbf{e}_i(t)\|^2|\mathcal{F}_t] = \zeta^2 \sum_{j=1}^n W_{ij}^2(t) \le \zeta^2$. Hence, the conditions of Assumption 1 are satisfied.

3. Auxiliary lemmas. The following lemmas are used to facilitate the proof of the main result of the paper. Lemma 3.1 follows from the Cauchy–Schwarz inequality and the geometric-arithmetic inequality, and its proof is omitted here for the sake of brevity. We present the proofs of Lemmas 3.2 and 3.5 in Appendix B.

LEMMA 3.1. *The inequality* $\|\boldsymbol{u}+\boldsymbol{v}\|^2 \le (1+\omega)\|\boldsymbol{u}\|^2 + (1+\omega^{-1})\|\boldsymbol{v}\|^2$ *holds for any pair of vectors* $\boldsymbol{u}$, $\boldsymbol{v}$, *and any scalar* $\omega > 0$. *Similarly, for matrices* U *and* V *and any scalar* $\omega > 0$, *their* $\mathbf{r}$*-norms satisfy* $\|U+V\|_{\mathbf{r}}^2 \le (1+\omega)\|U\|_{\mathbf{r}}^2 + (1+\omega^{-1})\|V\|_{\mathbf{r}}^2$.

LEMMA 3.2. *Let* $\{\beta(t)\}$ *be a sequence in* $[0,1]$ *and* $\lambda > 0$. *Then*

$$\sum_{s=1}^{t-1} \beta(s) \prod_{k=s+1}^{t-1} (1-\lambda\beta(k)) \le \frac{1}{\lambda}, \quad \sum_{t=s+1}^{T} \beta(t) \prod_{k=s+1}^{t-1} (1-\lambda\beta(k)) \le \frac{1}{\lambda}.$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 06/30/23 to 128.101.167.180 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

LEMMA 3.3. *For any $\delta \in \mathbb{R}$, $\tau \geq 0$, and $T \geq 1$, we have*

$$\sum_{t=1}^{T}(t+\tau)^{\delta} \leq \begin{cases} \frac{\tau^{1+\delta}}{|1+\delta|} & \text{if } \delta < -1, \\ \ln\left(\frac{T}{\tau}+1\right) & \text{if } \delta = -1, \\ \frac{2^{1+\delta}}{1+\delta}(T+\tau)^{1+\delta} & \text{if } \delta > -1. \end{cases} \tag{3.1}$$

LEMMA 3.4. *For any $A \in \mathbb{R}^{n\times m}$ and $B \in \mathbb{R}^{m\times q}$, we have $\|AB\|_{\mathbf{r}} \leq \|A\|_{\mathbf{r}}\|B\|_F$.*

As we discussed in Remark 2.5, we cannot use the conventional results (e.g., in [9, 24, 25, 23, 36, 35, 39, 34, 2]) to bound the norm of a vector after averaging with matrices with *diminishing* weights. The following lemma provides a new bounding technique, which we will use in the proof of the main result of this work.

LEMMA 3.5. *Let $\{W(t)\}$ satisfy the connectivity assumption, Assumption 2, with parameters (B,η), and let $\{A(t)\}$ be given by $A(t) = (1-\beta(t))I + \beta(t)W(t)$, where $\beta(t) \in (0,1]$ for all t and $\{\beta(t)\}$ is a nonincreasing sequence. Then, for any matrix $U \in \mathbb{R}^{n\times d}$, parameter $\lambda = \frac{\eta \mathbf{r}_{\min}}{2Bn^2}$, and all $t > s \geq 1$, we have*

$$\left\|\left(A(t-1)A(t-2)\cdots A(s+1) - \mathbf{1}\mathbf{r}^T\right)U\right\|_{\mathbf{r}}^2 \leq \kappa \prod_{k=s+1}^{t-1}(1-\lambda\beta(k))\,\|U\|_{\mathbf{r}}^2,$$

where $\kappa = (1-B\lambda\beta_0)^{-1}$ and $\beta_0 = \beta(1)$.

4. Proof of Theorem 2.1. For our analysis, first we obtain an expression for the average reduction of the objective function $f(\cdot)$ at the average of the states, i.e., $\bar{\mathbf{x}}(t) = \sum_{i=1}^{n} r_i \mathbf{x}_i(t) = \mathbf{r}^T X(t)$. Recall that $\mathbf{r}^T W(t) = \mathbf{r}^T$ for all $t \geq 1$. Hence, multiplying both sides of (2.7) by $\mathbf{r}^T$, we get

$$\bar{\mathbf{x}}(t+1) = \bar{\mathbf{x}}(t) + \beta(t)\mathbf{r}^T E(t) - \alpha(t)\beta(t)\mathbf{r}^T \nabla f(X(t)).$$

From Assumption 3(a) and Lemma 3.4 in [4] we can conclude

$$f(\bar{\mathbf{x}}(t+1)) - f(\bar{\mathbf{x}}(t)) - \langle \nabla f(\bar{\mathbf{x}}(t)), \bar{\mathbf{x}}(t+1) - \bar{\mathbf{x}}(t)\rangle \leq \frac{K}{2}\left\|\bar{\mathbf{x}}(t+1) - \bar{\mathbf{x}}(t)\right\|^2,$$

or equivalently,

$$\begin{aligned} f(\bar{\mathbf{x}}(t+1)) \leq f(\bar{\mathbf{x}}(t)) &+ \beta(t)\left\langle \nabla f(\bar{\mathbf{x}}(t)), \mathbf{r}^T E(t)\right\rangle - \alpha(t)\beta(t)\left\langle \nabla f(\bar{\mathbf{x}}(t)), \mathbf{r}^T \nabla f(X(t))\right\rangle \\ &+ \beta(t)^2 \frac{K}{2}\left\|\mathbf{r}^T E(t) - \alpha(t)\mathbf{r}^T \nabla f(X(t))\right\|^2. \end{aligned}$$

Since $\{X(t)\}$ is adapted to the filtration $\{\mathcal{F}_t\}$ and using Assumption 1, we arrive at

$$\begin{aligned} \mathbb{E}[f(\bar{\mathbf{x}}(t+1))|\mathcal{F}_t] \leq f(\bar{\mathbf{x}}(t)) &- \alpha(t)\beta(t)\left\langle \nabla f(\bar{\mathbf{x}}(t)), \mathbf{r}^T \nabla f(X(t))\right\rangle \\ &+ \beta^2(t)\frac{K}{2}\mathbb{E}\left[\left\|\mathbf{r}^T\left(E(t) - \alpha(t)\nabla f(X(t))\right)\right\|^2 |\mathcal{F}_t\right]. \end{aligned} \tag{4.1}$$

Using the identity $2\langle \mathbf{x}-\mathbf{y}\rangle = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - \|\mathbf{y}-\mathbf{x}\|^2$, we can write

$$\begin{aligned} \left\langle \nabla f\left(\bar{\mathbf{x}}(t)\right), \mathbf{r}^T \nabla f(X(t))\right\rangle = &\left\|\nabla f\left(\bar{\mathbf{x}}(t)\right)\right\|^2 + \left\|\mathbf{r}^T \nabla f(X(t))\right\|^2 \\ &- 2\left\|\mathbf{r}^T \nabla f(X(t)) - \nabla f\left(\bar{\mathbf{x}}(t)\right)\right\|^2. \end{aligned} \tag{4.2}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 06/30/23 to 128.101.167.180 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Moreover, we have

$$\begin{aligned}\|\mathbf{r}^T\nabla f(X(t))-\nabla f(\bar{\mathbf{x}}(t))\|^2 &= \left\|\sum_{i=1}^n r_i\left(\nabla f_i(\mathbf{x}_i(t))-\nabla f_i(\bar{\mathbf{x}}(t))\right)\right\|^2\\ &\le \sum_{i=1}^n r_i\left\|\nabla f_i(\mathbf{x}_i(t))-\nabla f_i(\bar{\mathbf{x}}(t))\right\|^2\\ &\le K^2\sum_{i=1}^n r_i\|\mathbf{x}_i(t)-\bar{\mathbf{x}}(t)\|^2 = K^2\left\|X(t)-\mathbf{1}\bar{\mathbf{x}}(t)\right\|_{\mathbf{r}}, \end{aligned} \tag{4.3}$$

where the first inequality holds as $\|\cdot\|^2$ is a convex function and $\mathbf{r}$ is a stochastic vector, and the second inequality follows from Assumption 3(a).

Next, we analyze the last term in (4.1). Note that Assumption 1 implies that $\mathbb{E}[E(t)|\mathcal{F}_t]=0$, which leads to

$$\mathbb{E}\left[\left\|\mathbf{r}^T\left(E(t)-\alpha(t)\nabla f(X(t))\right)\right\|^2|\mathcal{F}_t\right] = \mathbb{E}[\left\|\mathbf{r}^T E(t)\right\|^2|\mathcal{F}_t] + \left\|\alpha(t)\mathbf{r}^T\nabla f(X(t))\right\|^2. \tag{4.4}$$

For the first term in (4.4), we again exploit Assumption 1, which implies

$$\mathbb{E}\left[\left\|\mathbf{r}^T E(t)\right\|^2|\mathcal{F}_t\right] = \mathbf{r}^T\mathbb{E}[E(t)E(t)^T|\mathcal{F}_t]\mathbf{r} \le \mathbf{r}^T(\gamma\mathbf{1}\mathbf{1}^T)\mathbf{r}=\gamma, \tag{4.5}$$

where we used the fact that $\mathbf{r}^T\mathbf{1}=1$ and the inequality holds since

$$\left[\mathbb{E}[E(t)E(t)^T|\mathcal{F}_t]\right]_{ij} = \mathbb{E}[\mathbf{e}_i(t)\mathbf{e}_j(t)|\mathcal{F}_t] \le \sqrt{\mathbb{E}[\|\mathbf{e}_i(t)\|^2|\mathcal{F}_t]\mathbb{E}[\|\mathbf{e}_j(t)\|^2|\mathcal{F}_t]} \le \gamma$$

for all $i,j\in[n]$. Thus, the last term in (4.1) is upper bounded as

$$\mathbb{E}\left[\left\|\mathbf{r}^T\left(E(t)-\alpha(t)\nabla f(X(t))\right)\right\|^2|\mathcal{F}_t\right] \le \gamma+\alpha^2(t)\|\mathbf{r}^T\nabla f(X(t))\|^2. \tag{4.6}$$

Therefore, replacing (4.2), (4.3), and (4.6) in (4.1) we get

$$\begin{aligned}&\mathbb{E}[f(\bar{\mathbf{x}}(t+1))|\mathcal{F}_t]\\ &\le f(\bar{\mathbf{x}}(t)) - \alpha(t)\beta(t)\left\langle\nabla f(\bar{\mathbf{x}}(t)), \mathbf{r}^T\nabla f(X(t))\right\rangle\\ &\quad+\beta^2(t)\frac{K}{2}\mathbb{E}\left[\left\|\mathbf{r}^T\left[E(t)-\alpha(t)\nabla f(X(t))\right]\right\|^2|\mathcal{F}_t\right]\\ &\le f(\bar{\mathbf{x}}(t))-\frac{1}{2}\alpha(t)\beta(t)\left(\|\nabla f(\bar{\mathbf{x}}(t))\|^2+\left\|\mathbf{r}^T\nabla f(X(t))\right\|^2-2K^2\sum_{i=1}^n r_i\|\bar{\mathbf{x}}(t)-\mathbf{x}_i(t)\|^2\right)\\ &\quad+\beta^2(t)\frac{K}{2}\left(\gamma+\alpha^2(t)\left\|\mathbf{r}^T\nabla f(X(t))\right\|^2\right)\\ &= f(\bar{\mathbf{x}}(t)) - \frac{1}{2}\alpha(t)\beta(t)\|\nabla f(\bar{\mathbf{x}}(t))\|^2 - \frac{1}{2}\alpha(t)\beta(t)\left(1-\alpha(t)\beta(t)K\right)\left\|\mathbf{r}^T\nabla f(X(t))\right\|^2\\ &\quad+\alpha(t)\beta(t)K^2\left\|X(t)-\mathbf{1}\bar{\mathbf{x}}(t)\right\|_{\mathbf{r}}^2+\beta^2(t)\frac{K}{2}\gamma. \end{aligned}$$

Taking the expectation of both sides leads to

$$\begin{aligned}\mathbb{E}[f(\bar{\mathbf{x}}(t+1))] \le{}& \mathbb{E}[f(\bar{\mathbf{x}}(t))] - \frac{1}{2}\alpha(t)\beta(t)\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2]\\ &-\frac{1}{2}\alpha(t)\beta(t)\left(1-\alpha(t)\beta(t)K\right)\mathbb{E}[\left\|\mathbf{r}^T\nabla f(X(t))\right\|^2]\\ &+\alpha(t)\beta(t)K^2\mathbb{E}[\|X(t)-\mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2]+\beta^2(t)\frac{K}{2}\gamma. \end{aligned} \tag{4.7}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 06/30/23 to 128.101.167.180 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

4.1. State deviations from the average state: $\mathbb{E}[\|\boldsymbol{X(t)-1\bar{x}(t)}\|_{\mathbf{r}}^{\mathbf{2}}]$. Note that the dynamics in (2.7) can be viewed as the linear time-varying system

$$X(t+1)=A(t)X(t)+U(t) \tag{4.8}$$

with $A(t)=((1-\beta(t))I+\beta(t)W(t))$ and $U(t)=\beta(t)E(t)-\alpha(t)\beta(t)\nabla f(X(t))$. The solution of (4.8) is given by

$$X(t)=\sum_{s=1}^{t-1}\Phi(t:s)U(s)+\Phi(t:0)X(1), \tag{4.9}$$

where $\Phi(t:s)=A(t-1)\cdots A(s+1)$ with $\Phi(t:t-1)=I$ is the transition matrix of the linear system (4.8). For simplicity of notation, let us define

$$P(t:s):=\beta(s)(\Phi(t:s)-\mathbf{1}\mathbf{r}^T)=\beta(s)\left(A(t-1)\cdots A(s+1)-\mathbf{1}\mathbf{r}^T\right).$$

As a result of Lemma 3.5, we have $\|P(t:s)U\|_{\mathbf{r}}\leq\pi(t:s)\|U\|_{\mathbf{r}}$, where

$$\pi(t:s):=\beta(s)\kappa^{\frac{1}{2}}\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))^{\frac{1}{2}}. \tag{4.10}$$

Now consider the dynamic in (4.9). Assuming $X(1)=\mathbf{0}$, we get

$$X(t)=\sum_{s=1}^{t-1}\Phi(t:s)U(s). \tag{4.11}$$

Multiplying both sides of (4.11) from the left by $\mathbf{r}^T$ and using $\mathbf{r}^TA(t)=\mathbf{r}^T$, we get

$$\bar{\mathbf{x}}(t)=\mathbf{r}^TX(t)=\sum_{s=1}^{t-1}\mathbf{r}^T\Phi(t:s)U(s)=\sum_{s=1}^{t-1}\mathbf{r}^TU(s). \tag{4.12}$$

Then, subtracting (4.12) from (4.11), and plugging in the definition of $U(s)$ we have

$$\begin{aligned}X(t)-\mathbf{1}\bar{\mathbf{x}}(t)&=\sum_{s=1}^{t-1}\Phi(t:s)U(s)-\sum_{s=1}^{t-1}\mathbf{1}\mathbf{r}^TU(s)=\sum_{s=1}^{t-1}(\Phi(t:s)-\mathbf{1}\mathbf{r}^T)U(s)\\&=\sum_{s=1}^{t-1}\beta(s)(\Phi(t:s)-\mathbf{1}\mathbf{r}^T)\Big[E(s)-\alpha(s)\nabla f(X(s))\Big]\\&=\sum_{s=1}^{t-1}P(t:s)E(s)-\sum_{s=1}^{t-1}\alpha(s)P(t:s)\nabla f(X(s)).\end{aligned}$$

Therefore, using Lemma 3.1 with $\omega=1$, we get

$$\begin{aligned}\|X(t)-\mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2&=\left\|\sum_{s=1}^{t-1}P(t:s)E(s)-\sum_{s=1}^{t-1}\alpha(s)P(t:s)\nabla f(X(s))\right\|_{\mathbf{r}}^2\\&\leq 2\left\|\sum_{s=1}^{t-1}P(t:s)E(s)\right\|_{\mathbf{r}}^2+2\left\|\sum_{s=1}^{t-1}\alpha(s)P(t:s)\nabla f(X(s))\right\|_{\mathbf{r}}^2.\end{aligned}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 06/30/23 to 128.101.167.180 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

By expanding $\|\sum_{s=1}^{t-1} P(t:s)E(s)\|_{\mathbf{r}}^2$, we get

$$\|X(t)-\mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2 = 2\sum_{s=1}^{t-1}\|P(t:s)E(s)\|_{\mathbf{r}}^2 + 2\sum_{s\neq q}\langle P(t:s)E(s), P(t:q)E(q)\rangle$$
$$+2\left\|\sum_{s=1}^{t-1}\alpha(s)P(t:s)\nabla f(X(s))\right\|_{\mathbf{r}}^2. \tag{4.13}$$

Recall from Assumption 1 that $\mathbb{E}[E(q)|\mathcal{F}_q]=0$. Moreover, since $E(s)$ is measurable with respect to $\mathcal{F}_q$ for $q>s$, we have

$$\begin{aligned}\mathbb{E}[\langle P(t:s)E(s), P(t:q)E(q)\rangle] &= \mathbb{E}[\mathbb{E}\left[\langle P(t:s)E(s), P(t:q)E(q)\rangle\right]|\mathcal{F}_q]\\ &= \mathbb{E}[\langle P(t:s)E(s), P(t:q)\mathbb{E}\left[E(q)|\mathcal{F}_q\right]\rangle] = 0.\end{aligned}$$

Using a similar argument for $q<s$ and conditioning on $\mathcal{F}_s$, we conclude that the above relation holds for all $q\neq s$. Therefore, taking the expectation of both sides of (4.13) and noting that the average of the second and forth terms is zero, we get

$$\mathbb{E}[\|X(t)-\mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2] \leq 2\sum_{s=1}^{t-1}\mathbb{E}[\|P(t:s)E(s)\|_{\mathbf{r}}^2] + 2\mathbb{E}\left[\left\|\sum_{s=1}^{t-1}\alpha(s)P(t:s)\nabla f(X(s))\right\|_{\mathbf{r}}^2\right]. \tag{4.14}$$

We continue with bounding the first term in (4.14). First note that Assumption 1 implies

$$\mathbb{E}[\|E(s)\|_{\mathbf{r}}^2] = \mathbb{E}\left[\mathbb{E}\left[\|E(s)\|_{\mathbf{r}}^2|\mathcal{F}_s\right]\right] = \mathbb{E}\left[\sum_{i=1}^n r_i\mathbb{E}\left[\|\mathbf{e}_i(s)\|^2|\mathcal{F}_s\right]\right] \leq \mathbb{E}\left[\sum_{i=1}^n r_i\gamma\right] = \gamma. \tag{4.15}$$

This together with Lemma 3.5 leads to

$$\begin{aligned}\sum_{s=1}^{t-1}\mathbb{E}[\|P(t:s)E(s)\|_{\mathbf{r}}^2] &\leq \left[\sum_{s=1}^{t-1}\beta^2(s)\kappa\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))\mathbb{E}[\|E(s)\|_{\mathbf{r}}^2]\right]\\ &\leq \gamma\kappa\sum_{s=1}^{t-1}\left[\beta^2(s)\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))\right].\end{aligned} \tag{4.16}$$

Using the triangle inequality for $\|\cdot\|_{\mathbf{r}}$, we can bound the second term in (4.14) as

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{s=1}^{t-1}\alpha(s)P(t:s)\nabla f(X(s))\right\|_{\mathbf{r}}^2\right] &\leq \mathbb{E}\left[\left(\sum_{s=1}^{t-1}\|\alpha(s)P(t:s)\nabla f(X(s))\|_{\mathbf{r}}\right)^2\right]\\ &= \sum_{1\leq s,q\leq t-1}\mathbb{E}[\alpha(s)\left\|P(t:s)\nabla f(X(s))\right\|_{\mathbf{r}}\alpha(q)\left\|P(t:q)\nabla f(X(q))\right\|_{\mathbf{r}}].\end{aligned} \tag{4.17}$$

Using Lemma 3.5 and $2ab\leq a^2+b^2$, we can upper bound this expression as

$$\begin{aligned}&\sum_{1\leq s,q\leq t-1}\mathbb{E}[\alpha(s)\left\|P(t:s)\nabla f(X(s))\right\|_{\mathbf{r}}\alpha(q)\left\|P(t:q)\nabla f(X(q))\right\|_{\mathbf{r}}]\\ &\qquad\leq \sum_{1\leq s,q\leq t-1}\mathbb{E}[\alpha(s)\pi(t:s)\left\|\nabla f(X(s))\right\|_{\mathbf{r}}\alpha(q)\pi(t:q)\left\|\nabla f(X(q))\right\|_{\mathbf{r}}]\end{aligned}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

$$
\begin{aligned}
&= \sum_{1\le s,q\le t-1} \pi(t:s)\pi(t:q)\mathbb{E}[\alpha(s)\left\|\nabla f(X(s))\right\|_{\mathbf{r}}\alpha(q)\left\|\nabla f(X(q))\right\|_{\mathbf{r}}]\\
&\le \frac{1}{2}\sum_{1\le s,q\le t-1} \pi(t:s)\pi(t:q)\mathbb{E}[\alpha^2(s)\left\|\nabla f(X(s))\right\|_{\mathbf{r}}^2+\alpha^2(q)\left\|\nabla f(X(q))\right\|_{\mathbf{r}}^2]\\
&= \sum_{1\le s,q\le t-1} \pi(t:s)\pi(t:q)\mathbb{E}[\alpha^2(s)\left\|\nabla f(X(s))\right\|_{\mathbf{r}}^2]\\
&= \left(\sum_{q=1}^{t-1}\pi(t:q)\right)\left(\sum_{s=1}^{t-1}\alpha^2(s)\pi(t:s)\mathbb{E}[\|\nabla f(X(s))\|_{\mathbf{r}}^2]\right), \qquad (4.18)
\end{aligned}
$$

where $\pi(t:s)$ is given by (4.10). Using $\sqrt{1-x}\le 1-x/2$ and Lemma 3.2 we have

$$
\sum_{q=1}^{t-1}\pi(t:q)=\sum_{q=1}^{t-1}\left[\beta(q)\kappa^{\frac{1}{2}}\prod_{k=q+1}^{t-1}(1-\lambda\beta(k))^{\frac{1}{2}}\right]\le\sum_{q=1}^{t-1}\beta(q)\kappa^{\frac{1}{2}}\prod_{k=q+1}^{t-1}\left(1-\frac{\lambda}{2}\beta(k)\right)\le\frac{2}{\lambda}\kappa^{\frac{1}{2}}. \qquad (4.19)
$$

Using this inequality in (4.18), we get

$$
\mathbb{E}\left[\left\|\sum_{s=1}^{t-1}\alpha(s)P(t:s)\nabla f(X(s))\right\|_{\mathbf{r}}^2\right]\le\frac{2}{\lambda}\kappa^{\frac{1}{2}}\sum_{s=1}^{t-1}\left[\alpha^2(s)\pi(t:s)\mathbb{E}[\|\nabla f(X(s))\|_{\mathbf{r}}^2]\right]. \qquad (4.20)
$$

Finally, using the bounds obtained in (4.16) and (4.20) in (4.14) we arrive at

$$
\begin{aligned}
\mathbb{E}[\|X(t)-\mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2]&\le 2\gamma\kappa\sum_{s=1}^{t-1}\left[\beta^2(s)\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))\right]\\
&\quad+\frac{4}{\lambda}\kappa^{\frac{1}{2}}\sum_{s=1}^{t-1}\left[\alpha^2(s)\pi(t:s)\mathbb{E}[\|\nabla f(X(s))\|_{\mathbf{r}}^2]\right]. \qquad (4.21)
\end{aligned}
$$

4.2. Analysis of the overall deviation: $\sum_{t=1}^{T}\boldsymbol{\alpha(t)\beta(t)}\mathbb{E}[\|\boldsymbol{X(t)-1\bar{x}(t)}\|_{\mathbf{r}}^{\mathbf{2}}]$. Our goal here is to bound the overall weighted deviation of the states from their average. First recall the bound for $\mathbb{E}[\|X(t)-\bar{\mathbf{x}}(t)\mathbf{1}\|_{\mathbf{r}}^2]$, derived in section 4.1 for each t. Our goal here is to bound

$$
\begin{aligned}
\sum_{t=1}^{T}\alpha(t)\beta(t)\mathbb{E}[\|X(t)-\mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2]&\le 2\gamma\kappa\sum_{t=1}^{T}\alpha(t)\beta(t)\sum_{s=1}^{t-1}\left[\beta^2(s)\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))\right]\\
&\quad+\frac{4}{\lambda}\kappa^{\frac{1}{2}}\sum_{t=1}^{T}\alpha(t)\beta(t)\sum_{s=1}^{t-1}\left[\alpha^2(s)\pi(t:s)\mathbb{E}[\|\nabla f(X(s))\|_{\mathbf{r}}^2]\right]. \qquad (4.22)
\end{aligned}
$$

Focusing on the first term in (4.22), we can write

$$
\begin{aligned}
&\sum_{t=1}^{T}\left[\alpha(t)\beta(t)\sum_{s=1}^{t-1}\left[\beta^2(s)\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))\right]\right]\\
&=\sum_{s=1}^{T-1}\left[\beta^2(s)\sum_{t=s+1}^{T}\left[\alpha(t)\beta(t)\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))\right]\right]\\
&\le\sum_{s=1}^{T-1}\left[\alpha(s)\beta^2(s)\sum_{t=s+1}^{T}\left[\beta(t)\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))\right]\right]\le\frac{1}{\lambda}\sum_{s=1}^{T-1}\alpha(s)\beta^2(s), \qquad (4.23)
\end{aligned}
$$

Downloaded 06/30/23 to 128.101.167.180 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 06/30/23 to 128.101.167.180 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

where the first inequality is due to the fact that $\alpha(t) \leq \alpha(s)$ for $t > s$, and the second one follows from Lemma 3.2. Similarly, using the fact that $\alpha(t) \leq \alpha(s)$ for $t > s$, for the second term in (4.22), we have

$$\begin{aligned}
&\sum_{t=1}^{T}\left[\alpha(t)\beta(t)\sum_{s=1}^{t-1}\left[\alpha^2(s)\pi(t:s)\mathbb{E}[\|\nabla f(X(s))\|_{\mathbf{r}}^2]\right]\right]\\
&=\sum_{s=1}^{T-1}\left[\alpha^2(s)\mathbb{E}[\|\nabla f(X(s))\|_{\mathbf{r}}^2]\sum_{t=s+1}^{T}\alpha(t)\beta(t)\pi(t:s)\right]\\
&\leq\sum_{s=1}^{T-1}\left[\alpha^3(s)\mathbb{E}[\|\nabla f(X(s))\|_{\mathbf{r}}^2]\sum_{t=s+1}^{T}\beta(t)\pi(t:s)\right]. \qquad (4.24)
\end{aligned}$$

Since $\pi(t:s)=\beta(s)\kappa^{\frac{1}{2}}\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))^{\frac{1}{2}}$, using $\sqrt{1-x}\leq 1-x/2$, we have

$$\begin{aligned}
\sum_{t=s+1}^{T}\beta(t)\pi(t:s)&=\sum_{t=s+1}^{T}\left[\beta(t)\beta(s)\kappa^{\frac{1}{2}}\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))^{\frac{1}{2}}\right]\\
&\leq\beta(s)\kappa^{\frac{1}{2}}\sum_{t=s+1}^{T}\left[\beta(t)\prod_{k=s+1}^{t-1}\left(1-\frac{\lambda}{2}\beta(k)\right)\right]\leq\frac{2}{\lambda}\beta(s)\kappa^{\frac{1}{2}}, \qquad (4.25)
\end{aligned}$$

where the last inequality follows from Lemma 3.2. Then, (4.24) and (4.25) imply

$$\sum_{t=1}^{T}\left[\alpha(t)\beta(t)\sum_{s=1}^{t-1}\left[\alpha^2(s)\pi(t:s)\mathbb{E}[\|\nabla f(X(s))\|_{\mathbf{r}}^2]\right]\right]\leq\frac{2\kappa^{\frac{1}{2}}}{\lambda}\sum_{s=1}^{T-1}\left[\alpha^3(s)\beta(s)\mathbb{E}\|\nabla f(X(s))\|_{\mathbf{r}}^2]\right].$$

Therefore, plugging this and (4.23) into (4.22), we can conclude

$$\begin{aligned}
&\sum_{t=1}^{T}\alpha(t)\beta(t)\mathbb{E}[\|X(t)-\mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2]\\
&\leq\frac{2\gamma\kappa}{\lambda}\sum_{t=1}^{T}\alpha(t)\beta^2(t)+\frac{8\kappa}{\lambda^2}\sum_{t=1}^{T}\left[\alpha^3(t)\beta(t)\mathbb{E}[\|\nabla f(X(t))\|_{\mathbf{r}}^2]\right]. \qquad (4.26)
\end{aligned}$$

4.3. Bounding $\mathbb{E}[\|\nabla f(X(t))\|_{\mathbf{r}}^2]$. In this part, we study $\mathbb{E}[\|\nabla f(X(t)\|_{\mathbf{r}}]^2$ to provide an upper bound for it. Following [29], we can rewrite $\nabla f(X(t))$ as

$$\nabla f(X(t))=3\left[\frac{1}{3}\left(\nabla f(X(t))-\nabla f(\mathbf{1}\bar{\mathbf{x}}(t))+\frac{1}{3}\left(\nabla f(\mathbf{1}\bar{\mathbf{x}}(t))-\mathbf{1}\nabla f(\bar{\mathbf{x}}(t))\right)+\frac{1}{3}\mathbf{1}\nabla f(\bar{\mathbf{x}}(t))\right],$$

where $\nabla f(\bar{\mathbf{x}}(t)):=\sum_{i=1}^{n}\nabla f_i(\bar{\mathbf{x}}(t))$. Then, since $\|\cdot\|_{\mathbf{r}}^2$ is a convex function, we have

$$\begin{aligned}
\mathbb{E}[\|\nabla f(X(t)\|_{\mathbf{r}}^2]&\leq 3\mathbb{E}[\|\nabla f(X(t))-\nabla f(\mathbf{1}\bar{\mathbf{x}}(t))\|_{\mathbf{r}}^2]\\
&\quad+3\mathbb{E}[\|\nabla f(\mathbf{1}\bar{\mathbf{x}}(t))-\mathbf{1}\nabla f(\bar{\mathbf{x}}(t))\|_{\mathbf{r}}^2]+3\mathbb{E}[\|\mathbf{1}\nabla f(\bar{\mathbf{x}}(t))\|_{\mathbf{r}}^2]. \qquad (4.27)
\end{aligned}$$

Next, we bound each term in (4.27). Using (4.3), we can write

$$\begin{aligned}
\mathbb{E}\left[\|\nabla f(X(t))-\nabla f(\mathbf{1}\bar{\mathbf{x}}(t))\|_{\mathbf{r}}^2\right]&=\mathbb{E}\left[\sum_{i=1}^{n}r_i\|\nabla f_i(\mathbf{x}_i(t))-\nabla f_i(\bar{\mathbf{x}}(t))\|^2\right]\\
&\leq K^2\mathbb{E}\left[\|X(t)-\mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right]. \qquad (4.28)
\end{aligned}$$

Similarly, using the convexity of function $\|\cdot\|^2$, for the second term in (4.27) we have

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

$$
\begin{aligned}
&\mathbb{E}[\|\nabla f(\mathbf{1}\bar{\mathbf{x}}(t)) - \mathbf{1}\nabla f(\bar{\mathbf{x}}(t))\|_{\mathbf{r}}^2] = \mathbb{E}\left[\sum_{i=1}^{n} r_i \left\|\nabla f_i(\bar{\mathbf{x}}(t)) - \nabla f(\bar{\mathbf{x}}(t))\right\|^2\right] \\
&= \sum_{i=1}^{n} r_i \mathbb{E}\left[4\left\|\frac{1}{2}\Big(\nabla f_i(\bar{\mathbf{x}}(t)) - \nabla L(\bar{\mathbf{x}}(t))\Big) - \frac{1}{2}\Big(\nabla f(\bar{\mathbf{x}}(t)) - \nabla L(\bar{\mathbf{x}}(t))\Big)\right\|^2\right] \\
&\le \sum_{i=1}^{n} 2r_i \mathbb{E}[\|\nabla f_i(\bar{\mathbf{x}}(t)) - \nabla L(\bar{\mathbf{x}}(t))\|^2] + 2\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t)) - \nabla L(\bar{\mathbf{x}}(t))\|^2] \\
&\overset{(a)}{=} \sum_{i=1}^{n} 2r_i \mathbb{E}\left[\frac{1}{m_i^2}\left\|\sum_{j=1}^{m_i}\big[\nabla \ell(\bar{\mathbf{x}}(t), \xi_j^i) - \nabla L(\bar{\mathbf{x}}(t))\big]\right\|^2\right] \\
&\quad + 2\mathbb{E}\left[\frac{1}{N^2}\left\|\sum_{j=1}^{N}[\nabla \ell(\bar{\mathbf{x}}(t), \xi_j) - \nabla L(\bar{\mathbf{x}}(t))]\right\|^2\right] \\
&\overset{(b)}{=} \sum_{i=1}^{n} \frac{2r_i}{m_i^2}\sum_{j=1}^{m_i}\mathbb{E}[\|\nabla \ell(\bar{\mathbf{x}}(t), \xi_j^i) - \nabla L(\bar{\mathbf{x}}(t))\|^2] \\
&\quad + \frac{2}{N^2}\sum_{j=1}^{N}\mathbb{E}[\|\nabla \ell(\bar{\mathbf{x}}(t), \xi_j) - \nabla L(\bar{\mathbf{x}}(t))\|^2] \\
&\overset{(c)}{\le} \sum_{i=1}^{n} 2\frac{m_i}{N}\frac{1}{m_i^2} m_i \sigma^2 + \frac{2}{N^2} N\sigma^2 = \frac{2(n+1)}{N}\sigma^2, \qquad (4.29)
\end{aligned}
$$

where in (a) we replaced the definitions of $f_i(\bar{\mathbf{x}}(t))$ and $f(\bar{\mathbf{x}}(t))$ from (2.3) and (2.2), respectively, the equality in (b) holds since ξ_js are independent samples from the underlying distribution, and (c) follows from Assumption 3(b) and the fact that $r_i = m_i/N$ for $i \in [n]$. Finally, for the third term in (4.27), we have

$$
\mathbb{E}[\|\mathbf{1}\nabla f(\bar{\mathbf{x}}(t))\|_{\mathbf{r}}^2] = \mathbb{E}\left[\sum_{i=1}^{n} r_i \left\|\nabla f(\bar{\mathbf{x}}(t))\right\|^2\right] = \mathbb{E}\left[\left\|\nabla f(\bar{\mathbf{x}}(t))\right\|^2\right]. \qquad (4.30)
$$

Plugging (4.28)–(4.30) into (4.27), we get

$$
\mathbb{E}[\|\nabla f(X(t)\|_{\mathbf{r}}^2] \le 3K^2\mathbb{E}[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2] + \frac{6(n+1)}{N}\sigma^2 + 3\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2]. \qquad (4.31)
$$

Next, replacing this bound in (4.26), we arrive at

$$
\begin{aligned}
&\sum_{t=1}^{T}\alpha(t)\beta(t)\mathbb{E}\left[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right] \\
&\le \frac{2\gamma\kappa}{\lambda}\sum_{t=1}^{T}\alpha(t)\beta^2(t) + \frac{8\kappa}{\lambda^2}\sum_{t=1}^{T}\left[\alpha^3(t)\beta(t)\mathbb{E}[\|\nabla f(X(t))\|_{\mathbf{r}}^2]\right] \\
&\le \frac{2\gamma\kappa}{\lambda}\sum_{t=1}^{T}\alpha(t)\beta^2(t) + \frac{24\kappa K^2}{\lambda^2}\sum_{t=1}^{T}\alpha^3(t)\beta(t)\mathbb{E}[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2] \\
&\quad + \frac{48\kappa(n+1)\sigma^2}{N\lambda^2}\sum_{t=1}^{T}\alpha^3(t)\beta(t) + \frac{24\kappa}{\lambda^2}\sum_{t=1}^{T}\alpha^3(t)\beta(t)\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2]. \qquad (4.32)
\end{aligned}
$$

Now, define $\phi_{i,j}(T) := \sum_{t=1}^{T}\alpha^i(t)\beta^j(t)$. Then, $\frac{2\gamma\kappa}{\lambda}\sum_{t=1}^{T}\alpha(t)\beta^2(t) = \epsilon_1\phi_{1,2}(T)$ and $\frac{48\kappa(n+1)\sigma^2}{N\lambda^2}\sum_{t=1}^{T}\alpha^3(t)\beta(t) = \epsilon_2\phi_{3,1}(T)$, where $\epsilon_1 := \frac{2\gamma\kappa}{\lambda}$ and $\epsilon_2 := \frac{48\kappa(n+1)\sigma^2}{N\lambda^2}$.

Downloaded 06/30/23 to 128.101.167.180 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 06/30/23 to 128.101.167.180 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Furthermore, we set $T_0 := \lceil (\frac{14\alpha_0\kappa^{\frac{1}{2}}K}{\lambda})^{\frac{1}{\nu}} \rceil$ such that $\frac{24\kappa K^2}{\lambda^2}\alpha^2(T_0) \le \frac{24}{196} < \frac{1}{2}$. Then, for $T \ge T_0$ we can rewrite (4.32) as

$$\begin{aligned}
&\sum_{t=1}^{T_0} \alpha(t)\beta(t)\mathbb{E}[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2] + \sum_{t=T_0+1}^{T} \alpha(t)\beta(t)\mathbb{E}[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2] \\
&\le \epsilon_1\phi_{1,2}(T) + \epsilon_2\phi_{3,1}(T) + \frac{24\kappa}{\lambda^2}\sum_{t=1}^{T} \alpha^3(t)\beta(t)\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2] \\
&\quad + \frac{24\kappa K^2}{\lambda^2}\left[\sum_{t=1}^{T_0} \alpha^3(t)\beta(t)\mathbb{E}[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2] + \sum_{t=T_0+1}^{T} \alpha^3(t)\beta(t)\mathbb{E}[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2]\right] \\
&\le \epsilon_1\phi_{1,2}(T) + \epsilon_2\phi_{3,1}(T) + \frac{24\kappa}{\lambda^2}\sum_{t=1}^{T} \alpha^3(t)\beta(t)\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2] \\
&\quad + \frac{24\kappa K^2}{\lambda^2}\left[\sum_{t=1}^{T_0} \alpha^3(t)\beta(t)\mathbb{E}[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2]\right. \\
&\quad \left. + \alpha^2(T_0)\sum_{t=T_0+1}^{T} \alpha(t)\beta(t)\mathbb{E}[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2]\right] \\
&\le \epsilon_1\phi_{1,2}(T) + \epsilon_2\phi_{3,1}(T) + \frac{24\kappa}{\lambda^2}\sum_{t=1}^{T} \alpha^3(t)\beta(t)\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2] \\
&\quad + \epsilon_3 + \frac{1}{2}\sum_{t=T_0+1}^{T} \alpha(t)\beta(t)\mathbb{E}[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2],
\end{aligned}$$

where $\epsilon_3 := \frac{24\kappa K^2}{\lambda^2}\sum_{t=1}^{T_0} \alpha^3(t)\beta(t)\mathbb{E}[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2]$ does not grow with T, and the second inequality holds since $\alpha(t)$ is a nonincreasing sequence. Therefore, we have

$$\begin{aligned}
&\sum_{t=1}^{T} \alpha(t)\beta(t)\mathbb{E}[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2] \\
&\qquad \le 2\sum_{t=1}^{T_0} \alpha(t)\beta(t)\mathbb{E}[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2] + \sum_{t=T_0+1}^{T} \alpha(t)\beta(t)\mathbb{E}[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2] \\
&\qquad \le 2\epsilon_1\phi_{1,2}(T) + 2\epsilon_2\phi_{3,1}(T) + 2\epsilon_3 + \frac{48\kappa}{\lambda^2}\sum_{t=1}^{T} \alpha^3(t)\beta(t)\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2]. \qquad (4.33)
\end{aligned}$$

4.4. Back to the main dynamics. Recall the dynamics in (4.7), that is,

$$\begin{aligned}
&\mathbb{E}[f(\bar{\mathbf{x}}(t+1))] \le \mathbb{E}[f(\bar{\mathbf{x}}(t))] - \frac{1}{2}\alpha(t)\beta(t)\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2] + \frac{K\gamma}{2}\beta^2(t) \\
&- \frac{1}{2}\alpha(t)\beta(t)\left(1 - \alpha(t)\beta(t)K\right)\mathbb{E}[\left\|\mathbf{r}^T\nabla f(X(t))\right\|^2] + \alpha(t)\beta(t)K^2\mathbb{E}[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2].
\end{aligned}$$

Summing (4.7) for $t = 1, 2, \ldots, T$ and using (4.33) we get

$$\begin{aligned}
&\mathbb{E}[f(\bar{\mathbf{x}}(T+1))] \le \mathbb{E}[f(\bar{\mathbf{x}}(1))] - \frac{1}{2}\sum_{t=1}^{T} \alpha(t)\beta(t)\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2] + \frac{K\gamma}{2}\sum_{t=1}^{T} \beta^2(t) \\
&- \frac{1}{2}\sum_{t=1}^{T} \alpha(t)\beta(t)\left(1 - \alpha(t)\beta(t)K\right)\mathbb{E}[\left\|\mathbf{r}^T\nabla f(X(t))\right\|^2]
\end{aligned}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 06/30/23 to 128.101.167.180 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

$$
\begin{aligned}
&+K^2\sum_{t=1}^{T}\alpha(t)\beta(t)\mathbb{E}[\|X(t)-\mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2]\\
&\le \mathbb{E}[f(\bar{\mathbf{x}}(1))]-\frac{1}{2}\sum_{t=1}^{T}\alpha(t)\beta(t)\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2]+\frac{K}{2}\gamma\phi_{0,2}(T)\\
&\quad-\frac{1}{2}\sum_{t=1}^{T}\alpha(t)\beta(t)\left(1-\alpha(t)\beta(t)K\right)\mathbb{E}[\|\mathbf{r}^T\nabla f(X(t))\|^2]\\
&\quad+K^2\left[2\epsilon_1\phi_{1,2}(T)+2\epsilon_2\phi_{3,1}(T)+2\epsilon_3+\frac{48\kappa}{\lambda^2}\sum_{t=1}^{T}\alpha^3(t)\beta(t)\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2]\right], \qquad (4.34)
\end{aligned}
$$

where $\phi_{0,2}(T)=\sum_{t=1}^{T}\beta^2(t)$. Next, note that for $t\ge T_0=\lceil(\frac{14\alpha_0\kappa^{\frac{1}{2}}K}{\lambda})^{\frac{1}{\nu}}\rceil$ we have $\alpha(t)\beta(t)K\le\alpha(T_0)\beta(T_0)K\le\alpha(T_0)K\le\frac{\lambda}{14\kappa^{\frac{1}{2}}}<1$, where the last inequality holds since $\lambda\le 1$ and $\kappa>1$ (see Lemma 3.5). Therefore, the coefficient $1-\alpha(t)\beta(t)K$ is nonnegative for $t\ge T_0$. Thus, for $T\ge T_0$ we have

$$
\begin{aligned}
&\frac{1}{2}\sum_{t=1}^{T}\alpha(t)\beta(t)\left(1-\alpha(t)\beta(t)K\right)\mathbb{E}[\|\mathbf{r}^T\nabla f(X(t))\|^2]\\
&\quad\ge\frac{1}{2}\sum_{t=1}^{T_0}\alpha(t)\beta(t)\left(1-\alpha(t)\beta(t)K\right)\mathbb{E}[\|\mathbf{r}^T\nabla f(X(t))\|^2]:=\epsilon_4, \qquad (4.35)
\end{aligned}
$$

where ϵ_4 does not grow with T. Similarly, $\frac{48\kappa K^2}{\lambda^2}\alpha^2(T_0)\le\frac{48}{196}\alpha_0\le\frac{1}{4}$. Therefore, for $T>T_0$, the last summation in (4.34) can be upper bounded by

$$
\begin{aligned}
&\frac{48\kappa K^2}{\lambda^2}\sum_{t=1}^{T}\alpha^3(t)\beta(t)\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2]\\
&\quad=\frac{48\kappa K^2}{\lambda^2}\left[\sum_{t=1}^{T_0}\alpha^3(t)\beta(t)\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2]+\sum_{t=T_0+1}^{T}\alpha^3(t)\beta(t)\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2]\right],\\
&\quad\le\epsilon_5+\frac{48\kappa K^2}{\lambda^2}\alpha^2(T_0)\sum_{t=T_0+1}^{T}\alpha(t)\beta(t)\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2]\\
&\quad\le\epsilon_5+\frac{1}{4}\sum_{t=T_0+1}^{T}\alpha(t)\beta(t)\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2], \qquad (4.36)
\end{aligned}
$$

where $\epsilon_5:=\frac{48\kappa K^2}{\lambda}\sum_{t=1}^{T_0}\alpha^3(t)\beta(t)\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2]$ does not depend on T. Next, plugging (4.35) and (4.36) into (4.34), for $T>T_0$, we get

$$
\begin{aligned}
\mathbb{E}[f(\bar{\mathbf{x}}(T+1))]&\le\mathbb{E}[f(\bar{\mathbf{x}}(1))]+\frac{K}{2}\phi_{0,2}(T)+2K^2\left(\epsilon_1\phi_{1,2}(T)+\epsilon_2\phi_{3,1}(T)+\epsilon_3\right)\\
&\quad-\frac{1}{2}\sum_{t=1}^{T}\alpha(t)\beta(t)\left(1-\alpha(t)\beta(t)K\right)\mathbb{E}[\|\mathbf{r}^T\nabla f(X(t))\|^2]\\
&\quad-\frac{1}{2}\sum_{t=1}^{T}\alpha(t)\beta(t)\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2]+\frac{48\kappa K^2}{\lambda^2}\sum_{t=1}^{T}\alpha^3(t)\beta(t)\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2]\\
&\le\mathbb{E}[f(\bar{\mathbf{x}}(1))]+\frac{K}{2}\phi_{0,2}(T)+2K^2\left(\epsilon_1\phi_{1,2}(T)+\epsilon_2\phi_{3,1}(T)+\epsilon_3\right)-\epsilon_4
\end{aligned}
$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 06/30/23 to 128.101.167.180 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

$$
\begin{aligned}
&- \frac{1}{2}\sum_{t=1}^{T_0}\alpha(t)\beta(t)\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2] - \frac{1}{2}\sum_{t=T_0+1}^{T}\alpha(t)\beta(t)\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2] \\
&+ \epsilon_5 + \frac{1}{4}\sum_{t=T_0+1}^{T}\alpha(t)\beta(t)\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2] \\
&\le \mathbb{E}[f(\bar{\mathbf{x}}(1))] + \frac{K}{2}\phi_{0,2}(T) + 2K^2\left(\epsilon_1\phi_{1,2}(T) + \epsilon_2\phi_{3,1}(T) + \epsilon_3\right) - \epsilon_4 \\
&+ \epsilon_5 - \frac{1}{4}\sum_{t=1}^{T}\alpha(t)\beta(t)\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2].
\end{aligned}
$$

By rearrangement of the terms above, we get

$$
\begin{aligned}
&\sum_{t=1}^{T}\alpha(t)\beta(t)\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2] \\
&\quad \le 4\mathbb{E}[f(\bar{\mathbf{x}}(1))] - 4\mathbb{E}[f(\bar{\mathbf{x}}(T+1))] + 2K\phi_{0,2}(T) \\
&\qquad + 8K^2(\epsilon_1\phi_{1,2}(T) + \epsilon_2\phi_{3,1}(T) + \epsilon_3) - 4\epsilon_4 + 4\epsilon_5 \\
&\quad \le 4\mathbb{E}[f(\bar{\mathbf{x}}(1))] - 4\mathbb{E}[f(\mathbf{x}^\star)] + 2K\phi_{0,2}(T) \\
&\qquad + 8K^2(\epsilon_1\phi_{1,2}(T) + \epsilon_2\phi_{3,1}(T) + \epsilon_3) - 4\epsilon_4 + 4\epsilon_5 \\
&\quad = \epsilon_6 + 2K\phi_{0,2}(T) + 8K^2\epsilon_1\phi_{1,2}(T) + 8K^2\epsilon_2\phi_{3,1}(T) \\
&\quad \le \epsilon_6 + (2K + 8K^2\epsilon_1\alpha_0)\phi_{0,2}(T) + 8K^2\epsilon_2\phi_{3,1}(T), \qquad (4.37)
\end{aligned}
$$

where $\epsilon_6 := 8K^2\epsilon_3 - 4\epsilon_4 + 4\epsilon_5 + 4\mathbb{E}[f(\bar{\mathbf{x}}(1))] - 4\mathbb{E}[f(\mathbf{x}^\star)]$ is a constant (does not depend on T), and the last inequality in (4.37) follows from the fact that

$$
\phi_{1,2}(T) = \sum_{t=1}^{T}\alpha(t)\beta^2(t) = \alpha_0\sum_{t=1}^{T}\frac{1}{(t+\tau)^\nu}\beta^2(t) \le \alpha_0\sum_{t=1}^{T}\beta^2(t) = \alpha_0\phi_{0,2}(T). \qquad (4.38)
$$

4.5. Bound on the moments of $\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2]$ and $\mathbb{E}[\|\boldsymbol{X}(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2]$. The inequality in (4.37) provides us with an upper bound on the temporal average of $\{\alpha(t)\beta(t)\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2]\}$. However, our goal is to derive a bound on the temporal average of $\{\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2]\}$. To this end, define the convergence measure

$$
M_\theta(\nu,\mu) := \left[\frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2]\right)^\theta\right]^{\frac{1}{\theta}}
$$

for a given $\theta \in (0,1)$. Note that by Hölder's inequality [13, Theorem 6.2] for any $p, q > 1$ with $\frac{1}{p} + \frac{1}{q} = 1$, and nonnegative sequences $\{a_t\}_{t=1}^T$ and $\{b_t\}_{t=1}^T$, we have

$$
\left(\sum_{t=1}^{T}a_t b_t\right)^q \le \left(\sum_{t=1}^{T}a_t^p\right)^{\frac{q}{p}}\left(\sum_{t=1}^{T}b_t^q\right). \qquad (4.39)
$$

Let $(p,q) := (\frac{1}{1-\theta}, \frac{1}{\theta})$ so that $\frac{1}{p} + \frac{1}{q} = 1$. Furthermore, define

$$
a_t := \left(\frac{1}{\alpha(t)\beta(t)}\right)^\theta = \frac{(t+\tau)^{(\mu+\nu)\theta}}{(\alpha_0\beta_0)^\theta} \text{ and } b_t := \left(\alpha(t)\beta(t)\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2]\right)^\theta. \qquad (4.40)
$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 06/30/23 to 128.101.167.180 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Then, applying Hölder's inequality (4.39), we arrive at

$$M_\theta(\nu,\mu)=\left[\frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2]\right)^\theta\right]^{\frac{1}{\theta}}=\left(\frac{1}{T}\sum_{t=1}^{T}a_t b_t\right)^q\le\frac{1}{T^q}\left(\sum_{t=1}^{T}a_t^p\right)^{\frac{q}{p}}\left(\sum_{t=1}^{T}b_t^q\right). \tag{4.41}$$

It remains to upper bound the terms in the right-hand side (RHS) of (4.41). First, using Lemma 3.3 we get

$$\sum_{t=1}^{T}a_t^p=\frac{1}{(\alpha_0\beta_0)^{\frac{\theta}{1-\theta}}}\sum_{t=1}^{T}(t+\tau)^{\frac{(\nu+\mu)\theta}{1-\theta}}\le\frac{2^{1+\frac{(\nu+\mu)\theta}{1-\theta}}}{(\alpha_0\beta_0)^{\frac{\theta}{1-\theta}}\left(1+\frac{(\nu+\mu)\theta}{1-\theta}\right)}(T+\tau)^{1+\frac{(\nu+\mu)\theta}{1-\theta}}.$$

Therefore,

$$\left(\sum_{t=1}^{T}a_t^p\right)^{\frac{q}{p}}\le\frac{2^{\frac{1-\theta}{\theta}+(\nu+\mu)}}{\alpha_0\beta_0\left(1+\frac{(\nu+\mu)\theta}{1-\theta}\right)^{\frac{1-\theta}{\theta}}}(T+\tau)^{\frac{1-\theta}{\theta}+\nu+\mu}:=\epsilon_7(\theta)(T+\tau)^{\frac{1-\theta}{\theta}+\nu+\mu}. \tag{4.42}$$

Next, continuing from (4.37), for $T\ge T_0$ we can write

$$\begin{aligned}\sum_{t=1}^{T}b_t^q&=\sum_{t=1}^{T}\alpha(t)\beta(t)\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2]\le\epsilon_6+(2K+8K^2\epsilon_1\alpha_0)\phi_{0,2}(T)+8K^2\epsilon_2\phi_{3,1}(T)\\&\le\max\left\{3\epsilon_6,(6K+24K^2\epsilon_1\alpha_0)\phi_{0,2}(T),24K^2\epsilon_2\phi_{3,1}(T)\right\},\end{aligned} \tag{4.43}$$

where the last inequality follows from the fact that $a+b+c\le\max\{3a,3b,3c\}$. Then, we can use Lemma 3.3 and the fact that $2^{1-2\mu}\le 2$ for $0\le\mu<1$, to bound $\phi_{0,2}(T)$, as

$$\phi_{0,2}(T)=\sum_{t=1}^{T}\beta^2(t)=\frac{1}{\beta_0^2}\sum_{t=1}^{T}(t+\tau)^{-2\mu}\le\begin{cases}\frac{\tau^{1-2\mu}}{\beta_0^2|1-2\mu|} & \text{if } \mu>1/2,\\ \frac{1}{\beta_0^2}\ln\left(\frac{T}{\tau}+1\right) & \text{if } \mu=1/2,\\ \frac{2(T+\tau)^{1-2\mu}}{\beta_0^2(1-2\mu)} & \text{if } 0\le\mu<1/2.\end{cases} \tag{4.44}$$

Similarly, applying Lemma 3.3 on $\phi_{3,1}(T)$, we get

$$\phi_{3,1}(T)=\sum_{t=1}^{T}\alpha^3(t)\beta(t)=\frac{1}{\alpha_0^3\beta_0}\sum_{t=1}^{T}(t+\tau)^{-3\nu-\mu}\le\begin{cases}\frac{\tau^{1-3\nu-\mu}}{\alpha_0^3\beta_0|1-3\nu-\mu|} & \text{if } 3\nu+\mu>1,\\ \frac{1}{\alpha_0^3\beta_0}\ln\left(\frac{T}{\tau}+1\right) & \text{if } 3\nu+\mu=1,\\ \frac{2(T+\tau)^{1-3\nu-\mu}}{\alpha_0^3\beta_0(1-3\nu-\mu)} & \text{if } 0\le 3\nu+\mu<1,\end{cases} \tag{4.45}$$

in which $2^{1-3\nu-\mu}$ is bounded by 2, for $0\le 3\nu+\mu<1$. Note that the upper bound in (4.43) is the maximum of a constant and two $\phi_{\cdot,\cdot}(T)$ functions. Moreover, depending on the values of μ and ν, $\phi_{\cdot,\cdot}(T)$ functions can be upper bounded as in (4.44) and (4.45). Figure 2 illustrates the four regions of (μ,ν). In the following we first, analyze the interior of the four regions, and then study the boundary cases.

(Region I) $\mu>1/2$ and $3\nu+\mu>1$. Recall from (4.44) and (4.45) that $\phi_{0,2}(T)$ and $\phi_{3,1}(T)$ are both upper bounded by constants. Hence, in this regime, (4.43) leads to

$$\begin{aligned}\sum_{t=1}^{T}b_t^q&\le\max\left\{3\epsilon_6,(6K+24K^2\epsilon_1\alpha_0)\phi_{0,2}(T),24K^2\epsilon_2\phi_{3,1}(T)\right\}\\&\le\max\left\{3\epsilon_6,(6K+24K^2\epsilon_1\alpha_0)\frac{\tau^{1-2\mu}}{\beta_0^2|1-2\mu|},24K^2\epsilon_2\frac{\tau^{1-3\nu-\mu}}{\alpha_0^3\beta_0|1-3\nu-\mu|}\right\}:=\epsilon_8.\end{aligned} \tag{4.46}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

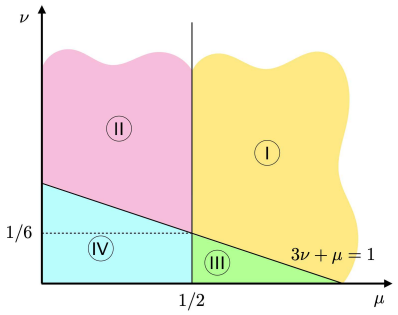

FIG. 2. *Regions of* (μ, ν).

Note that ϵ_8 is a constant. Plugging (4.42) and (4.46) into (4.41), we arrive at

$$M_\theta(\nu,\mu) \le \frac{1}{T^{1/\theta}}\epsilon_7(\theta)\cdot(T+\tau)^{\frac{1-\theta}{\theta}+\nu+\mu}\cdot\epsilon_8 = \mathcal{O}\left(T^{-(1-\nu-\mu)}\right).$$

(Region II) $\mu < 1/2$ and $3\nu+\mu > 1$. Note that (4.44) and (4.45) imply that $\phi_{0,2}(T)$ and $\phi_{3,1}(T)$ are upper bounded by a polynomial (in T) and a constant. Since for sufficiently large T, a polynomial function beats any constant, we can write

$$\sum_{t=1}^{T} b_t^q \le \frac{12K+48K^2\epsilon_1\alpha_0}{\beta_0^2(1-2\mu)}(T+\tau)^{1-2\mu} := \epsilon_9\cdot(T+\tau)^{1-2\mu}. \tag{4.47}$$

This together with (4.41) and (4.42) implies

$$M_\theta(\nu,\mu) \le \frac{1}{T^{1/\theta}}\epsilon_7(\theta)(T+\tau)^{\frac{1-\theta}{\theta}+\nu+\mu}\cdot\epsilon_9\cdot(T+\tau)^{1-2\mu} = \mathcal{O}\left(T^{-(\mu-\nu)}\right).$$

(Region III) $\mu > 1/2$ and $3\nu+\mu < 1$. Recall from (4.44) and (4.45) that $\phi_{0,2}(T)$ and $\phi_{3,1}(T)$ are upper bounded by a constant and a polynomial function of T, respectively. Therefore, for sufficiently large T, we get

$$\sum_{t=1}^{T} b_t^q \le \left(\frac{48K^2\epsilon_2}{\alpha_0^3\beta_0(1-3\nu-\mu)}\right)(T+\tau)^{1-3\nu-\mu} := \epsilon_{10}\cdot(T+\tau)^{1-3\nu-\mu}. \tag{4.48}$$

Therefore, plugging (4.48) and (4.42) into (4.41), we get

$$M_\theta(\nu,\mu) \le \frac{1}{T^{1/\theta}}\epsilon_7(\theta)(T+\tau)^{\frac{1-\theta}{\theta}+\nu+\mu}\cdot\epsilon_{10}\cdot(T+\tau)^{1-3\nu-\mu} = \mathcal{O}\left(T^{-2\nu}\right).$$

(Region IV) $\mu < 1/2$ and $3\nu+\mu < 1$. In this region, we can use (4.44) and (4.45) to upper bound both $\phi_{0,2}(T)$ and $\phi_{3,1}(T)$ by polynomial functions. Thus, for sufficiently large T, we get

$$\begin{aligned}\sum_{t=1}^{T} b_t^q &\le \max\left\{\frac{12K+48K^2\epsilon_1\alpha_0}{\beta_0^2(1-2\mu)}, \frac{48K^2\epsilon_2}{\alpha_0^3\beta_0(1-3\nu-\mu)}\right\}(T+\tau)^{\max\{1-2\mu,1-3\nu-\mu\}}\\ &:= \epsilon_{11}\cdot(T+\tau)^{\max\{1-2\mu,1-3\nu-\mu\}}.\end{aligned} \tag{4.49}$$

Then, we can plug (4.49) and (4.42) into (4.41) to conclude

$$\begin{aligned}M_\theta(\nu,\mu) &\le \frac{1}{T^{1/\theta}}\epsilon_7(\theta)(T+\tau)^{\frac{1-\theta}{\theta}+\nu+\mu}\cdot\epsilon_{11}\cdot(T+\tau)^{\max\{1-2\mu,1-3\nu-\mu\}}\\ &= \mathcal{O}\left(T^{-\min\{\mu-\nu,2\nu\}}\right).\end{aligned}$$

The result of the four cases above concludes the first claim of Theorem 2.1.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 06/30/23 to 128.101.167.180 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Downloaded 06/30/23 to 128.101.167.180 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Recall that our goal is to find (ν,μ) that achieve the best convergence rate. This is equivalent to optimizing the exponent of $1/T$ in each of the four (open) regions I$-$IV (shown in Figure 2). Interestingly, it turns out that the respective supremum value in all four regions is $1/3$, which corresponds to the boundary point $(\nu^\star,\mu^\star)=(1/6,1/2)$. However, this point does not belong to any of the corresponding open sets, which motivates the convergence analysis of M_θ for $(\nu^\star,\mu^\star)=(1/6,1/2)$.

(Boundary case) $\boldsymbol{\mu=1/2}$ and $\boldsymbol{3\nu+\mu=1}$. First note that the two lines of interest intersect at $(\nu^\star,\mu^\star)=(1/6,1/2)$, as shown in Figure 2. Applying Lemma 3.3 on $\phi_{0,2}(T)$ and $\phi_{3,1}(T)$ for $(\nu^\star,\mu^\star)=(1/6,1/2)$ we get $\phi_{0,2}(T)\le\frac{1}{\beta_0^2}\ln(\frac{T}{\tau}+1)$ and $\phi_{3,1}(T)\le\frac{1}{\alpha_0^3\beta_0}\ln(\frac{T}{\tau}+1)$. Therefore, (4.43) reduces to

$$\sum_{t=1}^{T} b_t^q \le \max\left\{\frac{6K+24K^2\epsilon_1\alpha_0}{\beta_0^2},\frac{24K^2\epsilon_2}{\alpha_0^3\beta_0}\right\}\ln\left(\frac{T}{\tau}+1\right):=\epsilon_{12}\cdot\ln\left(\frac{T}{\tau}+1\right). \tag{4.50}$$

Plugging (4.50) and (4.42) into (4.41), we arrive at

$$M_\theta\left(\frac{1}{6},\frac{1}{2}\right)\le\frac{1}{T^{1/\theta}}\epsilon_7(\theta)(T+\tau)^{\frac{1-\theta}{\theta}+\frac{2}{3}}\cdot\epsilon_{12}\cdot\ln\left(\frac{T}{\tau}+1\right)=\mathcal{O}\left(T^{-1/3}\ln T\right), \tag{4.51}$$

which is the second claim of the theorem.

5. Conclusion. We have studied nonconvex distributed optimization over time-varying networks with lossy information sharing. We proposed and studied a two-time-scale decentralized algorithm including a damping mechanism for incoming information from the neighboring agents as well as local cost functions' gradients. We presented the convergence rate for various choices of the diminishing step-size parameters. By optimizing the achieved rate over all feasible choices for parameters, the algorithm obtains a convergence rate of $\mathcal{O}(T^{-1/3+\epsilon})$ for nonconvex distributed optimization problems over time-varying networks for any $\epsilon>0$. Further work in this area may include the constrained/projection variation of this work.

Appendix A. Proof of Proposition 2.4. First note that we cannot directly conclude the proposition from Theorem 2.1, since the theorem only holds for $\theta\in(0,1)$ and not $\theta=1$. In order to show the claim, for a vector $\mathbf{y}\in\mathbb{R}^T$ and some $\theta\in(0,1)$ we define[1] $\|\mathbf{y}\|_\theta:=(|y_1|^\theta+|y_2|^\theta+\cdots+|y_T|^\theta)^{1/\theta}$. Then we have $\|\mathbf{y}\|_1\le\|\mathbf{y}\|_\theta$ since

$$\left(\frac{\|\mathbf{y}\|_\theta}{\|\mathbf{y}\|_1}\right)^\theta=\frac{|y_1|^\theta+\cdots+|y_T|^\theta}{(|y_1|+\cdots+|y_T|)^\theta}=\sum_{t=1}^{T}\left(\frac{|y_t|}{|y_1|+\cdots+|y_T|}\right)^\theta\ge\sum_{t=1}^{T}\frac{|y_t|}{|y_1|+\cdots+|y_T|}=1,$$

where the inequality holds since we have $0\le|y_t|/\sum_{i=1}^T|y_i|\le1$, and $\theta<1$.

Now, for the vector $\mathbf{y}$ with $y_t=\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2]$, we have

$$\begin{aligned}M_1&=\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2]=\frac{1}{T}\|\mathbf{y}\|_1\le\frac{1}{T}\|\mathbf{y}\|_\theta=\frac{1}{T}\left(\sum_{t=1}^{T}\left(\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2]\right)^\theta\right)^{1/\theta}\\&=\frac{1}{T^{1-1/\theta}}\left(\frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2]\right)^\theta\right)^{1/\theta}=\frac{1}{T^{1-1/\theta}}M_\theta.\end{aligned}$$

[1]Note that for $\|\mathbf{y}\|_\theta$ is not a norm, since it is not a subadditive function for $\theta<1$.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 06/30/23 to 128.101.167.180 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Then, from Theorem 2.1 for $(\nu^\star,\mu^\star)=(1/6,1/2)$ and $\theta=\frac{2}{2+\epsilon}$, we get

(A.1)
$$M_1 \le \frac{1}{T^{1-1/\theta}} M_\theta \le \frac{1}{T^{1-1/\theta}} \mathcal{O}\left(T^{-1/3}\ln T\right) = T^{\epsilon/2}\mathcal{O}\left(T^{-1/3}\ln T\right) = \mathcal{O}\left(T^{-1/3+\epsilon}\right),$$

where the last equality holds since $\ln T = \mathcal{O}(T^{\epsilon/2})$ for any $\epsilon>0$. Similarly, for the vector $\mathbf{z}\in\mathbb{R}^T$ with $z_t := \mathbb{E}[\|X(t)-\mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2]$ and any $\theta\in(0,1)$ we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|X(t)-\mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2] = \frac{1}{T}\|\mathbf{z}\|_1 \le \frac{1}{T}\|\mathbf{z}\|_\theta = \frac{1}{T}\left[\sum_{t=1}^{T}\left(\mathbb{E}[\|X(t)-\mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2]\right)^\theta\right]^{\frac{1}{\theta}}$$

(A.2)
$$= \frac{1}{T^{1-1/\theta}}\left[\frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}[\|X(t)-\mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2]\right)^\theta\right]^{\frac{1}{\theta}}.$$

We need to bound the RHS of (A.2). Let $a_t := (\alpha(t)\beta(t))^{-\theta} = (t+\tau)^{2\theta/3}/(\alpha_0\beta_0)^\theta$ and $c_t := (\alpha(t)\beta(t)\mathbb{E}[\|X(t)-\mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2])^\theta$. Then, using the Hölder's inequality in (4.39) for $(p,q)=(\frac{1}{1-\theta},\frac{1}{\theta})$ we can write

(A.3)
$$\left[\frac{1}{T}\sum_{t=1}^{T}\left(\mathbb{E}[\|X(t)-\mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2]\right)^\theta\right]^{\frac{1}{\theta}} = \left(\frac{1}{T}\sum_{t=1}^{T}a_t c_t\right)^q \le \frac{1}{T^q}\left(\sum_{t=1}^{T}a_t^p\right)^{\frac{q}{p}}\left(\sum_{t=1}^{T}c_t^q\right).$$

Note that $(\sum_{t=1}^{T}a_t^p)^{\frac{q}{p}}$ is bounded in (4.42). Moreover, for $\sum_{t=1}^{T}c_t^q$ we can write

(A.4)
$$\begin{aligned}
\sum_{t=1}^{T}c_t^q &= \sum_{t=1}^{T}\alpha(t)\beta(t)\mathbb{E}[\|X(t)-\mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2]\\
&\overset{(a)}{\le} 2\epsilon_1\phi_{1,2}(T)+2\epsilon_2\phi_{3,1}(T)+2\epsilon_3+\frac{48\kappa}{\lambda^2}\sum_{t=1}^{T}\alpha^3(t)\beta(t)\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2]\\
&\overset{(b)}{\le} 2\epsilon_1\phi_{1,2}(T)+2\epsilon_2\phi_{3,1}(T)+2\epsilon_3+\frac{48\kappa\alpha_0^2}{\lambda^2}\sum_{t=1}^{T}\alpha(t)\beta(t)\mathbb{E}[\|\nabla f(\bar{\mathbf{x}}(t))\|^2]\\
&\overset{(c)}{\le} 2\alpha_0\epsilon_1\phi_{0,2}(T)+2\epsilon_2\phi_{3,1}(T)+2\epsilon_3+\frac{48\kappa\alpha_0^2}{\lambda^2}\big(\epsilon_6+K(2+8K\alpha_0\epsilon_1)\phi_{0,2}(T)+8K^2\epsilon_2\phi_{3,1}(T)\big)\\
&\le 2\epsilon_3+\frac{48\kappa\alpha_0^2\epsilon_6}{\lambda^2}+\left(2\alpha_0\epsilon_1+\frac{96\kappa\alpha_0^2K(1+4K\alpha_0\epsilon_1)}{\lambda^2}\right)\phi_{0,2}(T)\\
&\quad+\left(2\epsilon_2+\frac{384\kappa\alpha_0^2K^2\epsilon_2}{\lambda^2}\right)\phi_{3,1}(T)\\
&\overset{(d)}{\le} 2\epsilon_3+\frac{48\kappa\alpha_0^2\epsilon_6}{\lambda^2}+\left(\frac{2\alpha_0\epsilon_1}{\beta_0^2}+\frac{96\kappa\alpha_0^2K(1+4K\alpha_0\epsilon_1)}{\lambda^2\beta_0^2}\right)\ln\left(\frac{T}{\tau}+1\right)\\
&\quad+\left(\frac{2\epsilon_2}{\alpha_0^3\beta_0}+\frac{384\kappa K^2\epsilon_2}{\lambda^2\alpha_0\beta_0}\right)\ln\left(\frac{T}{\tau}+1\right)\overset{(e)}{\le}\epsilon_{13}\cdot\ln\left(\frac{T}{\tau}+1\right),
\end{aligned}$$

where (a) follows from (4.33), (b) holds since $\alpha^2(t)=\frac{\alpha_0^2}{(t+\tau)^{1/3}}\le\alpha_0^2$, the inequality in (c) follows from (4.37) and (4.38), we have used a bounds in (4.44) and (4.45) for $(\nu^\star,\mu^\star)=(1/6,1/2)$ in (d), and (e) holds since the constant term $2\epsilon_3+48\kappa\alpha_0^2\epsilon_6/\lambda^2$ is

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 06/30/23 to 128.101.167.180 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

upper bounded by $\ln(T/\tau+1)$ for large enough T. Next, plugging (4.42) for $\nu^\star+\mu^\star = 2/3$ and (A.4) into (A.3) and (A.2) and setting $\theta = \frac{2}{2+\epsilon}$ we arrive at

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Big[\|\mathbf{x}_i(t)-\bar{\mathbf{x}}(t)\|^2\Big] \le \frac{1}{\mathbf{r}_{\min}}\left[\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|X(t)-\mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2]\right]$$

$$\le \frac{1}{\mathbf{r}_{\min}}\frac{1}{T^{1-1/\theta}}\frac{1}{T^{1/\theta}}\epsilon_7(\theta)(T+\tau)^{\frac{1-\theta}{\theta}+\frac{2}{3}}\cdot\epsilon_{13}\cdot\ln\left(\frac{T}{\tau}+1\right) = \mathcal{O}\left(T^{-1/3+\epsilon}\right), \tag{A.5}$$

where the last equality holds since $\ln T = \mathcal{O}(T^{\epsilon/2})$ for any $\epsilon>0$. Finally, combining (2.11) and (2.12) and using Assumption 3 and Lemma 3.1 (for $\omega=1$) we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Big[\|\nabla f(\mathbf{x}_i(t))\|^2\Big] \le \frac{2}{T}\sum_{t=1}^{T}\Big\{\mathbb{E}\Big[\|\nabla f(\mathbf{x}_i(t))-\nabla f(\bar{\mathbf{x}}(t))\|^2\Big]+\mathbb{E}\Big[\|\nabla f(\bar{\mathbf{x}}(t))\|^2\Big]\Big\}$$

$$\le \frac{2}{T}\sum_{t=1}^{T}\Big\{K^2\mathbb{E}\Big[\|\mathbf{x}_i(t)-\bar{\mathbf{x}}(t)\|^2\Big]+\mathbb{E}\Big[\|\nabla f(\bar{\mathbf{x}}(t))\|^2\Big]\Big\} \le \mathcal{O}(T^{-1/3+\epsilon})$$

for every $i\in[n]$. This concludes the proof of Proposition 2.4.

Appendix B. Proof of the auxiliary lemmas. In this section, we provide the proofs of the auxiliary lemmas presented in section 3.

Proof of Lemma 3.2. Let $g=\sum_{s=1}^{t-1}\beta(s)\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))$. Then, we have

$$\lambda g = \sum_{s=1}^{t-1}\lambda\beta(s)\prod_{k=s+1}^{t-1}(1-\lambda\beta(k)) = \sum_{s=1}^{t-1}(1-(1-\lambda\beta(s)))\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))$$

$$= \sum_{s=1}^{t-1}\left[\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))-\prod_{k=s}^{t-1}(1-\lambda\beta(k))\right] = 1-\prod_{k=1}^{t-1}(1-\lambda\beta(k)),$$

which follows from the fact that the corresponding sum is a telescopic sum. Dividing both sides by $\lambda>0$ and noticing that $\beta(k)\in[0,1]$, we arrive at the first desired inequality. To show the second inequality, we utilize the same idea for $h=\sum_{t=s+1}^{T}\beta(t)\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))$, where we have

$$\lambda h = \sum_{t=s+1}^{T}\left[\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))-\prod_{k=s+1}^{t}(1-\lambda\beta(k))\right]$$

$$= \prod_{k=s+1}^{s}(1-\lambda\beta(k))-\prod_{k=s+1}^{t}(1-\lambda\beta(k)) = 1-\prod_{k=s+1}^{T}(1-\lambda\beta(k)).$$

Again, dividing both sides by $\lambda>0$ and noting $\beta(k)\in[0,1]$ conclude the proof. ∎

Proof of Lemma 3.3. In order to prove the lemma, we separately analyze the cases $\delta<-1$, $\delta=-1$, $-1<\delta<0$, and $\delta\ge 0$. Note that for $\delta<0$, the function $h(t):=(t+\tau)^\delta$ is a decreasing function, and thus we have $\sum_{t=1}^{T}(t+\tau)^\delta \le \int_0^T(t+\tau)^\delta dt$. In the following we upper bound the latter integral for each regime of δ. When $\delta<-1$ we have

$$\sum_{t=1}^{T}(t+\tau)^\delta \le \int_0^T(t+\tau)^\delta dt = \frac{\tau^{1+\delta}-(T+\tau)^{1+\delta}}{-1-\delta} \le \frac{\tau^{1+\delta}}{-1-\delta}, \tag{B.1}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 06/30/23 to 128.101.167.180 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

which does not grow with T. For $\delta = -1$, we get

$$\sum_{t=1}^{T}(t+\tau)^{-1} \le \int_0^T (t+\tau)^{-1}dt = \ln(T+\tau) - \ln(\tau) = \ln\left(\frac{T}{\tau}+1\right). \tag{B.2}$$

When $-1 < \delta < 0$, we arrive at

$$\sum_{t=1}^{T}(t+\tau)^{\delta} \le \int_0^T (t+\tau)^{\delta}dt = \frac{(T+\tau)^{1+\delta}-\tau^{1+\delta}}{1+\delta} \le \frac{(T+\tau)^{1+\delta}}{1+\delta} \le \frac{2^{1+\delta}}{1+\delta}(T+\tau)^{1+\delta}, \tag{B.3}$$

where the last inequality follows from $2^{1+\delta} \ge 2^0 = 1$. Finally, for $\delta \ge 0$, the function $h(t)$ is an increasing function. Hence, we can write

$$\sum_{t=1}^{T}(t+\tau)^{\delta} \le \int_1^{T+1} (t+\tau)^{\delta}dt = \frac{(T+\tau+1)^{1+\delta}-(\tau+1)^{1+\delta}}{1+\delta} \le \frac{2^{1+\delta}}{1+\delta}(T+\tau)^{1+\delta}, \tag{B.4}$$

where the last step follows from $T+\tau+1 \ge 2(T+\tau)$ for any $T \ge 1$ and $\tau \ge 0$. ∎

Proof of Lemma 3.4. Recall that the (i,j)th entry of the matrix product AB is the inner product between the ith row of A and the jth column of B. Thus, using the Cauchy–Schwarz inequality, we have $|[AB]_{ij}| = |\langle A_i, B^j\rangle| \le \|A_i\|\|B^j\|$. Therefore,

$$\|[AB]_i\|^2 = \sum_{j=1}^{m}|[AB]_{ij}|^2 = \sum_{j=1}^{m}|\langle A_i, B^j\rangle|^2 \le \|A_i\|^2 \sum_{j=1}^{m}\|B^j\|^2 \le \|A_i\|^2 \|B\|_F^2.$$

Using this inequality and the definition of $\mathbf{r}$-norm, we get

$$\|AB\|_{\mathbf{r}}^2 = \sum_{i=1}^{n} r_i \|[AB]_i\|^2 \le \sum_{i=1}^{n} r_i \|A_i\|^2 \|B\|_F^2 = \|A\|_{\mathbf{r}}^2 \|B\|_F^2. \qquad ∎$$

Proof of Lemma 3.5. Due to the separable nature of $\|\cdot\|_{\mathbf{r}}$, i.e., $\|U\|_{\mathbf{r}}^2 = \sum_{j=1}^{d}\|U^j\|_{\mathbf{r}}^2$, without loss of generality, we may assume that $d = 1$. Thus, $U = \boldsymbol{u} \in \mathbb{R}^{n\times 1} = \mathbb{R}^n$ is a column vector. Define $V_{\mathbf{r}} : \mathbb{R}^n \to \mathbb{R}^+$ by $V_{\mathbf{r}}(\boldsymbol{u}) := \|\boldsymbol{u} - \mathbf{1}\mathbf{r}^T\boldsymbol{u}\|_{\mathbf{r}}^2 = \sum_{i=1}^{n} r_i(u_i - \mathbf{r}^T\boldsymbol{u})^2$. For notational simplicity, let $\boldsymbol{u}(s) = \boldsymbol{u} = (u_1, u_2, \dots, u_n)^T$, and $\boldsymbol{u}(k+1) = A(k+1)\boldsymbol{u}(k)$. Also with a slight abuse of notation, we denote $V_{\mathbf{r}}(\boldsymbol{u}(k))$ by $V_{\mathbf{r}}(k)$ for $k = s, \dots, t$. Using Lemma 3 in [36], we have

$$V_{\mathbf{r}}(t) = V_{\mathbf{r}}(s) - \sum_{k=s+1}^{t}\sum_{i<j} H_{ij}(k)(u_i(k) - u_j(k))^2, \tag{B.5}$$

where $H(k) = A^T(k)\mathrm{diag}(\mathbf{r})A(k)$. Note that $A(k)$ is a nonnegative matrix, and hence we have $H(k) \ge \mathbf{r}_{\min}A^T(k)A(k)$ for $k = s+1, \dots, t$. Also, note that since $A(k) = (1-\beta(k))I + \beta(k)W(k)$, then Assumption 2(b) implies that the minimum nonzero elements of $A(k)$ are bounded below by $\eta\beta(k)$. Therefore, since $\beta(k)$ is nonincreasing, on the window $k = s+1, \dots, s+B$, the minimum nonzero elements of $A(k)$ for k in this window are lower bounded by $\eta\beta(s+B)$. Without loss of generality, assume that the entries of $\boldsymbol{u}$ are sorted, i.e., $u_1 \le \cdots \le u_n$; otherwise, we can relabel the agents

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 06/30/23 to 128.101.167.180 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

(rows and columns of $A(k)$s and $\boldsymbol{u}$ to achieve this). Therefore, by Lemma 8 in [24], for (B.5), we have

$$\begin{aligned} V_{\mathbf{r}}(s+B) &\le V_{\mathbf{r}}(s) - \mathbf{r}_{\min} \sum_{k=s+1}^{s+B} \sum_{i<j} [A^T(k)A(k)]_{ij}(u_i(k)-u_j(k))^2 \\ &\le V_{\mathbf{r}}(s) - \frac{\eta \mathbf{r}_{\min}}{2}\beta(s+B)\sum_{\ell=1}^{n-1}(u_{\ell+1}-u_\ell)^2. \end{aligned} \tag{B.6}$$

We may comment here that although Lemma 8 in [24] is written for doubly stochastic matrices, and its statement is about the decrease of $V_{\mathbf{r}}(\mathbf{x})$ for the special case of $\mathbf{r}=\frac{1}{n}\mathbf{1}$, in fact it is a result on bounding $\sum_{k=s+1}^{s+B}\sum_{i<j}[A^T(k)A(k)]_{ij}(u_i(k)-u_j(k))^2$ for a sequence of B-connected stochastic matrices $A(k)$ in terms of the minimum nonzero entries of stochastic matrices $A(s+1),\dots,A(s+B)$.

Next, we use an argument similar to that in the proof of Theorem 18 in [24] to show that $\sum_{\ell=1}^{n-1}(u_{\ell+1}-u_\ell)^2 \ge n^{-2}V_{\mathbf{r}}(\boldsymbol{u})$. For $\boldsymbol{v}\in\mathbb{R}^n$ with $V_{\mathbf{r}}(\boldsymbol{v})>0$, define the quotient

$$h(\boldsymbol{v}) = \frac{\sum_{\ell=1}^{n-1}(v_{\ell+1}-v_\ell)^2}{\sum_{i=1}^{n} r_i(v_i-\mathbf{r}^T\boldsymbol{v})^2} = \frac{\sum_{\ell=1}^{n-1}(v_{\ell+1}-v_\ell)^2}{V_{\mathbf{r}}(\boldsymbol{v})}. \tag{B.7}$$

Note that $h(\boldsymbol{v})$ is invariant under scaling and translations by the all-one vector, i.e., $h(\omega\boldsymbol{v})=h(\boldsymbol{v})$ for all nonzero $\omega\in\mathbb{R}$ and $h(\boldsymbol{v}+\omega\mathbf{1})=h(\boldsymbol{v})$ for all $\omega\in\mathbb{R}$. Therefore,

$$\min_{\substack{v_1\le v_2\le\cdots\le v_n\\ V_{\mathbf{r}}(\boldsymbol{v})\neq 0}} h(\boldsymbol{v}) = \min_{\substack{v_1\le v_2\le\cdots\le v_n\\ \mathbf{r}^T\boldsymbol{v}=0, V_{\mathbf{r}}(\boldsymbol{v})=1}} h(\boldsymbol{v}) = \min_{\substack{v_1\le v_2\le\cdots\le v_n\\ \mathbf{r}^T\boldsymbol{v}=0, V_{\mathbf{r}}(\boldsymbol{v})=1}} \sum_{\ell=1}^{n-1}(v_{\ell+1}-v_\ell)^2. \tag{B.8}$$

The facts that $\mathbf{r}$ is a stochastic vector, $v_1\le\cdots\le v_n$, and $\mathbf{r}^T\boldsymbol{v}=0$ imply $v_1\le\mathbf{r}^T\boldsymbol{v}=0\le v_n$. Moreover, from $V_{\mathbf{r}}(\boldsymbol{v})=\sum_{i=1}^n r_i v_i^2=1$ we can conclude $\max(|v_1|,|v_n|)\ge 1/\sqrt{n}$. Let us consider the difference sequence $\hat{v}_\ell = v_{\ell+1}-v_\ell$ for $\ell=1,\dots,n-1$, for which we have $\sum_{i=1}^{n-1}\hat{v}_\ell = v_n - v_1 \ge v_n \ge \frac{1}{\sqrt{n}}$. Therefore, for the optimization problem (B.8) we have

$$\min_{\substack{v_1\le v_2\le\cdots\le v_n\\ V_{\mathbf{r}}(\boldsymbol{v})\neq 0}} h(\boldsymbol{v}) = \min_{\substack{v_1\le v_2\le\cdots\le v_n\\ \mathbf{r}^T\boldsymbol{v}=0, V_{\mathbf{r}}(\boldsymbol{v})=1}} \sum_{\ell=1}^{n-1}(v_{\ell+1}-v_\ell)^2 \ge \min_{\substack{\hat{v}_1,\dots,\hat{v}_{n-1}\ge 0\\ \sum_{i=1}^{n-1}\hat{v}_i\ge\frac{1}{\sqrt{n}}}} \sum_{\ell=1}^{n-1}\hat{v}_\ell^2 \overset{\text{(a)}}{\ge} \frac{1}{n(n-1)} \ge \frac{1}{n^2},$$

where the inequality in (a) follows from the Cauchy–Schwarz inequality which implies $(\sum_{\ell=1}^{n-1}\hat{v}_\ell^2)(\sum_{\ell=1}^{n-1}1^2)\ge(\sum_{\ell=1}^{n-1}\hat{v}_\ell)^2\ge\frac{1}{n}$. Therefore, we have $\sum_{\ell=1}^{n-1}(v_{\ell+1}-v_\ell)^2\ge n^{-2}V_{\mathbf{r}}(\boldsymbol{v})$ for $v_1\le\cdots\le v_n$ (note that this inequality also holds for $\boldsymbol{v}\in\mathbb{R}^n$ with $V_{\mathbf{r}}(\boldsymbol{v})=0$). Using this fact in (B.6) implies

$$V_{\mathbf{r}}(s+B) \le \Big(1-\frac{\eta\mathbf{r}_{\min}}{2n^2}\beta(s+B)\Big)V_{\mathbf{r}}(s). \tag{B.9}$$

Applying (B.9) for $\Delta := \lfloor\frac{t-1-s}{B}\rfloor$ steps recursively, we get

$$V_{\mathbf{r}}(s+\Delta B) \le \prod_{j=1}^{\Delta}\Big(1-\frac{\eta\mathbf{r}_{\min}}{2n^2}\beta(s+jB)\Big)V_{\mathbf{r}}(s).$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 06/30/23 to 128.101.167.180 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Then, since $(1-x)^{1/B} \le 1 - x/B$ and $\{\beta(k)\}$ is a nonincreasing sequence, we have

$$\begin{aligned}1 - \frac{\eta \mathbf{r}_{\min}}{2n^2}\beta(s+jB) &= \prod_{\ell=1}^{B}\Big(1 - \frac{\eta \mathbf{r}_{\min}}{2n^2}\beta(s+jB)\Big)^{1/B} \\ &\le \prod_{\ell=1}^{B}\Big(1 - \frac{\eta \mathbf{r}_{\min}}{2Bn^2}\beta(s+jB)\Big) \le \prod_{\ell=1}^{B}\left(1 - \lambda\beta(s+jB+\ell)\right).\end{aligned}$$

Since $V_{\mathbf{r}}(\cdot)$ is nonincreasing (see (B.5)), for $s + \Delta B \le t-1 < s + (\Delta+1)B$ we have

$$\begin{aligned}V_{\mathbf{r}}(t-1) \le V_{\mathbf{r}}(s+\Delta B) &\le \prod_{j=1}^{\Delta}\Big(1 - \frac{\eta \mathbf{r}_{\min}}{2n^2}\beta(s+jB)\Big) V_{\mathbf{r}}(s) \\ &\le \prod_{j=1}^{\Delta}\prod_{\ell=1}^{B}(1-\lambda\beta(s+jB+\ell))\, V_{\mathbf{r}}(s) \\ &= \prod_{k=s+B+1}^{s+(\Delta+1)B}(1-\lambda\beta(k))\, V_{\mathbf{r}}(s) \le \prod_{k=s+B+1}^{t-1}(1-\lambda\beta(k))\, V_{\mathbf{r}}(s). \qquad \text{(B.10)}\end{aligned}$$

Next, since $\{\beta(k)\}$ is a nonincreasing sequence, we have $\beta(k) \le \beta(1) = \beta_0$. Thus,

$$\prod_{k=s+1}^{s+B}(1-\lambda\beta(k)) \ge \prod_{k=s+1}^{s+B}(1-\lambda\beta_0) = (1-\lambda\beta_0)^B \ge 1 - B\lambda\beta_0. \qquad \text{(B.11)}$$

Therefore, combining (B.10) and (B.11), we get

$$\begin{aligned}V_{\mathbf{r}}(t-1) &\le \prod_{k=s+B+1}^{t-1}(1-\lambda\beta(k))\, V_{\mathbf{r}}(s) \le \frac{\prod_{k=s+1}^{s+B}(1-\lambda\beta(k))}{1-B\lambda\beta_0}\prod_{k=s+B+1}^{t-1}(1-\lambda\beta(k))\, V_{\mathbf{r}}(s) \\ &= \frac{1}{1-B\lambda\beta_0}\prod_{k=s+1}^{t-1}(1-\lambda\beta(k)) V_{\mathbf{r}}(s). \qquad \text{(B.12)}\end{aligned}$$

Now, we define $\Phi(t:s) = A(t-1)\cdots A(s+1)$ for $t \ge s$ with $\Phi(t:t-1) = I$. Note that Assumption 2(a) and $A(k) = (1-\beta(k))I + \beta(k)W(k)$ imply $\mathbf{r}^T\Phi(t:s) = \mathbf{r}^T$. Then, setting $\boldsymbol{u}(s) = \boldsymbol{u} = U$ and $\boldsymbol{u}(t-1) = \Phi(t:s)\boldsymbol{u}(s) = \Phi(t:s)U$, we can write

$$\begin{aligned}\big(\Phi(t:s) - \mathbf{1}\mathbf{r}^T\big)U &= \Phi(t:s)U - \mathbf{1}\mathbf{r}^T U = \Phi(t:s)U - \mathbf{1}\mathbf{r}^T\Phi(t:s)U \\ &= \boldsymbol{u}(t-1) - \mathbf{1}\mathbf{r}^T\boldsymbol{u}(t-1).\end{aligned}$$

Therefore, using (B.12), we conclude the desired result by noticing

$$\begin{aligned}\big\|\big(\Phi(t:s) - \mathbf{1}\mathbf{r}^T\big)U\big\|_{\mathbf{r}}^2 &= \big\|\boldsymbol{u}(t-1) - \mathbf{1}\mathbf{r}^T\boldsymbol{u}(t-1)\big\|_{\mathbf{r}}^2 = V_{\mathbf{r}}(t-1) \\ &\le \frac{1}{1-B\lambda\beta_0}\prod_{k=s+1}^{t-1}(1-\lambda\beta(k)) V_{\mathbf{r}}(s) \\ &= \frac{1}{1-B\lambda\beta_0}\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))\big\|\boldsymbol{u} - \mathbf{1}\mathbf{r}^T\boldsymbol{u}\big\|_{\mathbf{r}}^2 \\ &\le \frac{1}{1-B\lambda\beta_0}\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))\|\boldsymbol{u}\|_{\mathbf{r}}^2 = \frac{1}{1-B\lambda\beta_0}\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))\|U\|_{\mathbf{r}}^2, \qquad \text{(B.13)}\end{aligned}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

where the second inequality follows from $\|\boldsymbol{u} - \mathbf{1}\mathbf{r}^T\boldsymbol{u}\|_{\mathbf{r}}^2 + \|\mathbf{1}\mathbf{r}^T\boldsymbol{u}\|_{\mathbf{r}}^2 = \|\boldsymbol{u}\|_{\mathbf{r}}^2$. Applying (B.13) on each column of a matrix U, we get the same result for matrices. ∎

REFERENCES

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., *Tensorflow: A system for large-scale machine learning*, in Proceedings of the 12th USENIX Symposium on OSDI, 2016, pp. 265–283.

[2] A. Aghajan and B. Touri, *Distributed optimization over dependent random networks*, IEEE Trans. Automat. Control, (2022).

[3] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, *QSGD: Communication-efficient SGD via gradient quantization and encoding*, in Proceedings of NeurIPS, 2017, pp. 1709–1720.

[4] S. Bubeck, *Convex optimization: Algorithms and complexity*, Found. Trends Mach. Learn., 8 (2015), pp. 231–357.

[5] K. Cai and H. Ishii, *Quantized consensus and averaging on gossip digraphs*, IEEE Trans. Automat. Control, 56 (2011), pp. 2087–2100.

[6] J. Chen and A. H. Sayed, *Diffusion adaptation strategies for distributed optimization and learning over networks*, IEEE Trans. Signal Process., 60 (2012), pp. 4289–4305.

[7] T. Chen, Q. Ling, and G. B. Giannakis, *An online convex optimization approach to proactive network resource allocation*, IEEE Trans. Signal Process., 65 (2017), pp. 6350–6364.

[8] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, M. Ranzato, A. Senior, P. Tucker, et al., *Large scale distributed deep networks*, in Proceedings of NeurIPS, 2012, pp. 1223–1231.

[9] P. M. DeMarzo, D. Vayanos, and J. Zwiebel, *Persuasion bias, social influence, and unidimensional opinions*, Q. J. Econ., 118 (2003), pp. 909–968.

[10] P. Di Lorenzo and G. Scutari, *NEXT: In-network nonconvex optimization*, IEEE Trans. Signal Inform. Process. Netw., 2 (2016), pp. 120–136.

[11] T. T. Doan, *Nonlinear two-time-scale stochastic approximation convergence and finite-time performance*, IEEE Trans. Automat. Control, to appear, doi:10.1109/TAC.2022.3210147.

[12] R. Durrett, *Probability: Theory and Examples*, Cambridge University Press, Cambridge, UK, 2010.

[13] G. B. Folland, *Real Analysis: Modern Techniques and their Applications*, Pure Appl. Math. 40, John Wiley & Sons, New York, 1999.

[14] D. Jakovetić, J. Xavier, and J. M. Moura, *Fast distributed gradient methods*, IEEE Trans. Automat. Control, 59 (2014), pp. 1131–1146.

[15] P. H. Jin, Q. Yuan, F. Iandola, and K. Keutzer, *How to scale distributed deep learning?*, in NIPS Workshop MLSystems, 2016.

[16] S. Kar and J. M. Moura, *Distributed consensus algorithms in sensor networks with imperfect communication: Link failures and channel noise*, IEEE Trans. Signal Process., 57 (2008), pp. 355–369.

[17] A. Kashyap, T. Başar, and R. Srikant, *Quantized consensus*, Automatica, 43 (2007), pp. 1192–1203.

[18] A. Koloskova, T. Lin, S. U. Stich, and M. Jaggi, *Decentralized deep learning with arbitrary communication compression*, in Proceedings of ICLR, 2020.

[19] A. Koloskova, S. U. Stich, and M. Jaggi, *Decentralized stochastic optimization and gossip algorithms with compressed communication*, in Proceedings of ICML, 2019, pp. 3478–3487.

[20] V. R. Konda and J. N. Tsitsiklis, *Convergence rate of linear two-time-scale stochastic approximation*, Ann. Appl. Probab., 14 (2004), pp. 796–819.

[21] 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*, NeurIPS, 30 (2017).

[22] I. Lobel and A. Ozdaglar, *Distributed subgradient methods for convex optimization over random networks*, IEEE Trans. Automat. Control, 56 (2010), pp. 1291–1306.

[23] A. Nedić and A. Olshevsky, *Distributed optimization of strongly convex functions on directed time-varying graphs*, in Proceedings of the IEEE Global Conference on Signal and Information Processing, 2013, pp. 329–332.

[24] A. Nedić, A. Olshevsky, A. Ozdaglar, and J. N. Tsitsiklis, *On distributed averaging algorithms and quantization effects*, IEEE Trans. Automat. Control, 54 (2009), pp. 2506–2517.

[25] A. Nedić and A. Ozdaglar, *Distributed subgradient methods for multi-agent optimization*, IEEE Trans. Automat. Control, 54 (2009), pp. 48–61.

Downloaded 06/30/23 to 128.101.167.180 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

[26] A. Nedić, A. Ozdaglar, and P. A. Parrilo, *Constrained consensus and optimization in multi-agent networks*, IEEE Trans. Automat. Control, 55 (2010), pp. 922–938.
[27] Y. Pu, M. N. Zeilinger, and C. N. Jones, *Quantization design for distributed optimization*, IEEE Trans. Automat. Control, 62 (2016), pp. 2107–2120.
[28] M. Rabbat and R. Nowak, *Distributed optimization in sensor networks*, in Proceedings of IPSN, 2004, pp. 20–27.
[29] A. Reisizadeh, H. Taheri, A. Mokhtari, H. Hassani, and R. Pedarsani, *Robust and communication-efficient collaborative learning*, in Proceedings of NeurIPS, 2019, pp. 8388–8399.
[30] H. Reisizadeh, B. Touri, and S. Mohajer, *Distributed optimization over time-varying graphs with imperfect sharing of information*, IEEE Trans. Automat. Control, 2022, doi: 10.1109/TAC.2022.3207866.
[31] A. Ribeiro, *Ergodic stochastic optimization algorithms for wireless communication and networking*, IEEE Trans. Signal Process., 58 (2010), pp. 6369–6386.
[32] G. Scutari and Y. Sun, *Distributed nonconvex constrained optimization over time-varying digraphs*, Math. Program., 176 (2019), pp. 497–544.
[33] K. Srivastava and A. Nedic, *Distributed asynchronous constrained stochastic optimization*, IEEE J. Selected Topics Signal Process., 5 (2011), pp. 772–790.
[34] T. Tatarenko and B. Touri, *Non-convex distributed optimization*, IEEE Trans. Automat. Control, 62 (2017), pp. 3744–3757.
[35] B. Touri and C. Langbort, *On endogenous random consensus and averaging dynamics*, IEEE Trans. Control Network Syst., 1 (2014), pp. 241–248.
[36] B. Touri and A. Nedić, *Product of random stochastic matrices*, IEEE Trans. Automat. Control, 59 (2013), pp. 437–448.
[37] M. M. Vasconcelos, T. T. Doan, and U. Mitra, *Improved convergence rate for a distributed two-time-scale gradient method under random quantization*, in Proceedings of the Conference on Decision and Control, IEEE, 2021, pp. 3117–3122.
[38] J. Xu, S. Zhu, Y. C. Soh, and L. Xie, *Convergence of asynchronous distributed gradient methods over stochastic networks*, IEEE Trans. Automat. Control, 63 (2017), pp. 434–448.
[39] K. Yuan, Q. Ling, and W. Yin, *On the convergence of decentralized gradient descent*, SIAM J. Optim., 26 (2016), pp. 1835–1854.
[40] J. Zeng and W. Yin, *On nonconvex decentralized gradient descent*, IEEE Trans. Signal Process., 66 (2018), pp. 2834–2848.

Downloaded 06/30/23 to 128.101.167.180 . Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/terms-privacy

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.