Distributed Optimization Over Time-Varying Graphs With Imperfect Sharing of Information

Hadi Reisizadeh, Behrouz Touri, *Senior Member, IEEE*, and Soheil Mohajer, *Member, IEEE*

***Abstract*—We study strongly convex distributed optimization problems where a set of agents are interested in solving a separable optimization problem collaboratively. In this article, we propose and study a two-time-scale decentralized gradient descent algorithm for a broad class of lossy sharing of information over time-varying graphs. One time-scale fades out the (lossy) incoming information from neighboring agents, and one time-scale regulates the local loss functions' gradients. We show that assuming a proper choice of step-size sequences, certain connectivity conditions, and bounded gradients along the trajectory of the dynamics, the agents' estimates converge to the optimal solution with the rate of $\mathcal{O}(T^{-1/2})$. We also provide novel tools to study distributed optimization with diminishing averaging weights over time-varying graphs.**

***Index Terms*—Convex optimization, distributed multiagent system, distributed optimization, gradient descent algorithms, time-varying graphs.**

I. Introduction

EMERGENCE of Big Data analytics, modern computer architectures, storage, and data collection have led to a growing interest in the study of multiagent networked systems. These systems arise in various applications, such as sensor networks [1], [2], network routing [3], large-scale machine learning [4], power control [5], and distributed network resource allocations [6], [7], for which decentralized solutions offer promising results. In general and in the absence of a central entity, we are often dealing with a time-varying network of agents, each can perform local and on-device computation. The information can be shared throughout the network via local communication between neighboring agents. This communication among agents, especially when the dimension of the data is large, accounts for a significant delay in the overall running time of the algorithm. In this article, we study distributed optimization under those lossy and imperfect information sharing scenarios, and propose and analyze a gradient-based distributed algorithm that guarantees convergence to the optimum solution, in the presence of communication constraints.

Related Works: Various methods have been proposed and studied to solve distributed optimization problems in convex settings [8], [9], [10], [11], [12], [13], [14], [15], strongly convex settings [13], [16], [17], and nonconvex settings [18], [19]. For the convex objective functions, a subgradient method with a fixed step-size was proposed over time-varying graphs in [20]. It is shown that the objective cost function reduces at rates of $\mathcal{O}(T^{-1})$ until it reaches a neighbor of a minimizer of the original problem. To achieve exact convergence to a minimizer, various diminishing step-size subgradient methods have been proposed and studied [10], [12], [18], [19], [21]. Considering convex loss functions that are Lipschitz continuous and have bounded gradients, a subgradient-push algorithm was proposed in [21]. There it was shown that the objective cost function convergences at the rate of $\mathcal{O}(T^{-1/2}\ln T)$ over uniformly strongly connected, directed time-varying graphs. Under the same assumption and strong convexity for loss functions, a better rate $\mathcal{O}(T^{-1}\ln T)$ for the objective loss function plus squared consensus residual was shown in [21].

Almost all the aforementioned works on this domain consider distributed optimization with perfect sharing of information, i.e., the agents are allowed to communicate real-valued vectors perfectly over perfect communication channels. However, exchanging exact information among nodes initiates a massive communication overhead on the system that considerably slows down the convergence rate of these algorithms in real-world applications. Thus, it is reasonable to assume that each agent has access to a lossy version of neighboring agents' information.

To address lossy/noisy sharing of information, a (fixed step-size) decentralized gradient descent method was proposed in [22]. Assuming a fixed communication network and strongly convex local cost functions, there it was shown that for a given iteration T, the algorithm's parameters (depending on T) can be chosen such that the local estimate of each agent at iteration T is (roughly) within $c(T^{-1/2+\epsilon})$-distance of the global optimal solution for some $c>0$ and any $\epsilon>0$. Furthermore, the result holds for a termination time T that is required to satisfy $T\geq T_{\min}$, where $T_{\min}$ depends on ϵ as well as nonlocal parameters of the underlying fixed graph. Specifically, as ϵ goes to zero, $T_{\min}$ diverges to infinity. In a closely related recent work [23], a two-time-scale gradient descent algorithm has been presented for strongly convex loss functions. Assuming a *fixed* topology for the underlying network, uniform weighting of the local cost functions, and a specific scheme for lossy sharing of information, it is shown that the expected objective loss function achieves a rate of $\mathcal{O}(T^{-1/2}(\ln T)^2)$. In another related work [24], a two-time-scale gradient descent algorithm was presented for distributed constrained and convex optimization problems over an independent identically distributed (i.i.d.) communication graph with noisy communication links, and subgradient errors. It is shown that under certain conditions on the i.i.d. communication graph and proper choices of time-scale parameters, the proposed dynamics results in almost sure convergence of local states to the optimal point.

Contributions: We study distributed optimization problem for a *general* class of lossy/noisy information sharing over *time-varying* communication networks. The learning method relies only on local computations and received imperfect information from neighboring agents. We show that a two-time-scale gradient descent algorithm, with a proper choice of parameters, reaches the global optima (in L_2 and hence, in probability) for *every* agent with a rate of $\mathcal{O}(T^{-1/2})$. To achieve this, we make limiting assumptions including weight matrices

Manuscript received 23 February 2022; revised 18 August 2022; accepted 1 September 2022. Date of publication 19 September 2022; date of current version 28 June 2023. The work of H. Reisizadeh and S. Mohajer was supported by the National Science Foundation under Grant CCF-1749981. Recommended by Associate Editor A. Olshevsky. *(Corresponding author: Behrouz Touri.)*

Hadi Reisizadeh and Soheil Mohajer are with the University of Minnesota, Minneapolis, MN 55455 USA (e-mail: hadir@umn.edu; soheil@umn.edu).

Behrouz Touri is with the University of California San Diego, La Jolla, CA 92093 USA (e-mail: btouri@ucsd.edu).

Digital Object Identifier 10.1109/TAC.2022.3207866

admitting the same stationary distribution, B-connected communication network, and expected boundedness of the gradients along the trajectories of the dynamics. Although weight matrices admitting the same stationary distribution is a limiting assumption, but a stronger assumption, namely having doubly stochastic weight matrices, is commonly assumed in the distributed optimization literature [10], [13], [20], [22]. For the bounded gradient assumption, using a slightly more sophisticated proof in the extended version of this work [25], we show that the Lipschitz gradients result in expected bounded gradients for strongly convex functions.

In addition, in the existing works on distributed optimization [9], [12], [13], [17], [18], [20], [21], [22], [23] (with perfect or imperfect sharing of information), either the underlying communication network is assumed to be fixed, or the nonzero elements of the averaging weights are assumed to be uniformly bounded away from zero. In our proposed method, however, the weights are not uniformly bounded away from zero and they are evolving over an underlying time-varying communication network. One of the key contributions of this article is to develop tools and techniques to deal with diminishing averaging weights for distributed optimization over time-varying networks.

Notation: We denote the set of integers $\{1, 2, \ldots, n\}$ by $[n]$ and the set of nonnegative real numbers by $\mathbb{R}^+$. 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. For a row vector $\mathbf{x} \in \mathbb{R}^d$, we use $\|\mathbf{x}\|$ to denote the 2-norm of $\mathbf{x}$. For a positive vector $\mathbf{r} \in \mathbb{R}^n$ with $\sum_{i=1}^n r_i = 1$, the $\mathbf{r}$-norm of an $n \times d$ matrix A is defined 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. We denote the Frobenius norm of A by $\|A\|_F$, where $\|A\|_F^2 = \sum_{i=1}^n \sum_{j=1}^d |A_{ij}|^2$. Moreover, $A \geq B$ indicates that all the entries of $A - B$ are nonnegative. Finally, a nonnegative matrix $A \in \mathbb{R}^{n\times d}$ is called stochastic if $\sum_{j=1}^d A_{ij} = 1$ for every $i \in [n]$.

II. Problem Setup and Main Result

In this section, we discuss the problem formulation and the main result of this work.

A. Problem Setup

Consider a set of $n \geq 2$ agents that are connected through a time-varying network. Each agent $i \in [n]$ has access to a local cost function $f_i : \mathbb{R}^d \to \mathbb{R}$. The goal of this article is to minimize the function $f(\mathbf{x}) := \sum_{i=1}^n r_i f_i(\mathbf{x})$, or equivalently solve

$$\min_{\mathbf{x}_1,\ldots,\mathbf{x}_n\in\mathbb{R}^d} \sum_{i=1}^n r_i f_i(\mathbf{x}_i) \quad \text{s.t.} \quad \mathbf{x}_1 = \mathbf{x}_2 = \cdots = \mathbf{x}_n \tag{1}$$

where vector $\mathbf{r} = (r_1, r_2, \ldots, r_n)$ is a stochastic vector, i.e., $r_i \geq 0$ and $\sum_{i=1}^n r_i = 1$.

We represent the time-varying topology 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 time t, agent i can only send messages to its (out-)neighbors in $\mathcal{E}(t)$, i.e., all $j \in [n]$ such that $(i, j) \in \mathcal{E}(t)$. To achieve a consensus, the sequence $\{\mathcal{G}(t)\}$ should satisfy some desirable long-term connectivity properties, which will be discussed in Assumption 2.

To present our algorithm for solving (1) collaboratively, let us first discuss the general framework for lossy/noisy sharing of information that is considered in this work. We assume that each agent maintains the sate $\mathbf{x}_i(t) \in \mathbb{R}^d$, which is an estimate of the optimizer of (1), and has access to its local cost function's gradient information. Moreover, it has access to an imperfect weighted average of its neighbors states at time t, denoted by $\hat{\mathbf{x}}_i(t)$. More precisely, agent i has access to $\hat{\mathbf{x}}_i(t) = \sum_{j=1}^n W_{ij}(t)\mathbf{x}_j(t) + \mathbf{e}_i(t)$, where $W(t) = [W_{ij}(t)]$ is a row-stochastic matrix that is consistent with the underlying network $\mathcal{G}(t)$ (i.e., $W_{ij}(t) > 0$ if only if $(j, i) \in \mathcal{G}(t)$) and $\mathbf{e}_i(t)$ is a random noise vector in $\mathbb{R}^d$. Although this setting might look contrived, many practical multiagent communication settings satisfy this structural assumption. Later, we discuss two of such practical settings.

Now, we present our *Diminishing Mixing* (DIMIX) algorithm. In this algorithm, each agent i updates its current estimate by computing a diminishing convex combination of its own state and received noisy average estimate $\hat{\mathbf{x}}_i(t)$, moving along its local gradient. Mathematically, at each node $i \in [n]$, the update rule is given by

$$\mathbf{x}_i(t+1) = (1-\beta(t))\mathbf{x}_i(t) + \beta(t)\hat{\mathbf{x}}_i(t) - \alpha(t)\beta(t)\nabla f_i(\mathbf{x}_i(t)) \tag{2}$$

where $\alpha(t) = \frac{\alpha_0}{t^\nu}$, $\beta(t) = \frac{\beta_0}{t^\mu}$, and $\mu, \nu \in (0, 1)$ are the diminishing step-sizes of the algorithm. A similar dynamics is independently proposed and discussed for a particular subsetting (i.e., specific lossy sharing mechanism, weight vector $\mathbf{r}$, and a specific choice of $\nu, \mu > 0$) of our framework for time-invarying networks in [23]. For simplicity of notation, let

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

Using this notation, we can rewrite the update algorithm (2) in the compact matrix format as

$$X(t+1) = ((1-\beta(t))I + \beta(t)W(t))X(t) + \beta(t)E(t) - \alpha(t)\beta(t)\nabla f(X(t)). \tag{3}$$

B. Assumptions

Here, we discuss the assumptions on the agent i's neighbor average state estimate $\hat{\mathbf{x}}_i(t)$, the stochastic weight matrix $\{W(t)\}$, and local objective functions f_i, that we will use in the subsequent discussions.

Assumption 1 (Noise Assumption): We assume that the noise sequence $\{\mathbf{e}_i(t)\}$ satisfies

$$\mathbb{E}\left[\mathbf{e}_i(t) \mid \mathcal{F}_t\right] = 0 \text{ and } \mathbb{E}\left[\|\mathbf{e}_i(t)\|^2 \mid \mathcal{F}_t\right] \leq \gamma$$

for some $\gamma > 0$, all $i \in [n]$, and all $t \geq 1$. Here, $\{\mathcal{F}_t\}$ is the natural filtration for the process $\{X(t)\}$.

As mentioned before, to provide guarantees on the working of our algorithm, certain connectivity assumptions need to be satisfied among the agents over time.

Assumption 2 (Connectivity Assumption): We assume that the weight matrix sequence $\{W(t)\}$ satisfies the following.

1) *Stochastic with common stationary distribution:* $W(t)$ is nonnegative and $W(t)\mathbf{1} = \mathbf{1}$ and $\mathbf{r}^T W(t) = \mathbf{r}^T$ for all $t \geq 1$, where $\mathbf{1} \in \mathbb{R}^n$ is the all-one vector, and $\mathbf{r} > 0$ is the stochastic weight vector appearing in (1).
2) *Bounded nonzero elements:* There exists some $\eta > 0$ such that if for some $i, j \in [n]$ and $t \geq 1$ we have $W_{ij}(t) > 0$, then $W_{ij}(t) \geq \eta$.
3) *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 $\mathcal{E}(k) = \{(j, i) \mid W_{ij}(k) > 0\}$.

The following assumptions hold for the objective functions.

Assumption 3 (Function Assumptions): We assume the following properties on the function f_i for all i.

1) The function f_i is L-smooth, i.e., for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have that $\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$.

2) The function f_i is ρ-strongly convex, i.e., for any $\mathbf{x},\mathbf{y}\in\mathbb{R}^d$, we have $\langle\nabla f_i(\mathbf{x})-\nabla f_i(\mathbf{y}),\mathbf{x}-\mathbf{y}\rangle\geq\rho\|\mathbf{x}-\mathbf{y}\|^2$.
3) f_i has uniformly bounded gradients along the trajectory of the dynamics (3) $\{\mathbf{x}_i(t)\}$, i.e., there exists a scalar $K>0$ such that for all $t\geq 1$ and all $i\in[n]$

$$\|\nabla f_i(\mathbf{x}_i(t))\|^2\leq K,\quad \mathbf{x}_i(t)\in\mathbb{R}^d. \tag{4}$$

Remark 1: The properties of f_is in Assumption 3 can be immediately translated to similar properties for $f(\mathbf{x})=\sum_{i=1}^n r_i f_i(\mathbf{x})$. More precisely, the function f is also L-smooth and ρ-strongly convex, and has a bounded gradient with $\|\nabla f(X(t))\|_{\mathbf{r}}^2\leq K$.

C. Main Result and Discussion

The main result of this article is the following theorem.

Theorem 1: If the conditions in Assumptions 1–3 are satisfied for $\alpha(t)=\frac{\alpha_0}{t^\nu}$ and $\beta(t)=\frac{\beta_0}{t^\mu}$, and $\mu,\nu\in(0,1)$ when $\mu+\nu<1$, the dynamics (3) satisfy

$$\mathbb{E}\left[\|X(T)-\mathbf{1}\mathbf{x}^\star\|_{\mathbf{r}}^2\right]\leq\xi_1 T^{-\min(\mu,2\nu)}+\xi_2\exp\left(-\xi_3 T^{1-\mu-\nu}\right)+\xi_4 T^{-\min(\mu-\nu,2\nu)} \tag{5}$$

for any iteration $T\geq\max(T_1,T_2,T_3,T_4)$, where T_1, T_2, T_3, and T_4 are given in (16), (21), (30), and (39), respectively, and $\mathbf{x}^\star:=\arg\min f(\mathbf{x})$. Moreover, the constants ξ_ℓ for $\ell\in\{1,\dots,4\}$ are evaluated in (42). Furthermore, under the same assumptions, when $\mu+\nu=1$ and $\alpha_0\beta_0\geq\frac{\rho+L}{\rho L}\min(2\mu-1,2\nu)$, the dynamics (3) satisfy

$$\mathbb{E}\left[\|X(T)-\mathbf{1}\mathbf{x}^\star\|_{\mathbf{r}}^2\right]\leq\xi_1 T^{-\min(\mu,2\nu)}+\xi_5 T^{-\min(\mu-\nu,2\nu)} \tag{6}$$

for any iteration $T\geq\max(T_1,T_2,T_3)$, where the constant ξ_5 is determined in (43).

We refer to Section IV for the proof of Theorem 1.

Remark 2: Theorem 1 guarantees the exact convergence (in L_2 sense) of each local state to the global optimal with diminishing step-size even though the noises induced by random quantization and gradients are not vanishing with iterations. To maximize the exponents in the upper bounds (5) and (6), it can be verified that the solution is $\mu=3/4$ and $\nu=1/4$. Replacing this in (6), we conclude

$$\mathbb{E}\left[\|X(T)-\mathbf{1}\mathbf{x}^\star\|_{\mathbf{r}}^2\right]\leq\xi T^{-1/2}$$

for any $T\geq\max(T_1,T_2,T_3)=\max((2/\lambda\beta_0)^4,\alpha_0\beta_0(\rho+L)/2)$ and some constant $\xi>0$. Our algorithm and the main result are inspired by the fixed step-size variation of (3) proposed in [22] under the limited setting of time-invarying networks, uniform weights $\mathbf{r}$, and a particular choice of lossy sharing of information. In that setting, it is shown that for any *given stopping time* $T\geq T_{\min}$ and any $\epsilon>0$, the constant step-sizes $\alpha_0,\beta_0>0$ can be set such that

$$\mathbb{E}\left[\|X(T)-\mathbf{1}\mathbf{x}^\star\|_{\mathbf{r}}^2\right]\leq cT^{-1/2+\epsilon}$$

where c and $T_{\min}$ are positive constants depending on the problem's parameters (note that this is established for a fixed T). However, $T_{\min}\to\infty$ as $\epsilon\to 0$ [22]. Here, we provide a rigorous convergence rate analysis that reduces to $\mathcal{O}(T^{-1/2})$ for *every* iteration T. In Theorem 1, for the case $\mu+\nu=1$, the minimum number of required iterations is finite.

Examples for Stochastic Noisy State Estimation: The noisy estimation of the neighbors' state used in (2) may appear in various practical problem settings. In the following, we describe two of such scenarios.

Example 1: A practical scenario where the noisy average neighbor estimate arises is when we deal with noisy communication channels between the agents. Consider a wireless medium, in which the communication links between the agents are Gaussian channels, i.e., when node j sends its state $\mathbf{x}_j(t)$ to its neighbor i, the received signal at node i is $\mathbf{x}_j(t)+\mathbf{z}_{i,j}(t)$, where $\mathbf{z}_{i,j}(t)$ is a zero-mean Gaussian noise with variance σ^2, independent across (i,j), and t. Then, we have $\hat{\mathbf{x}}_i(t)=\sum_{j=1}^n W_{ij}(t)(\mathbf{x}_j(t)+\mathbf{z}_{i,j}(t))$. Therefore, in this case, $\hat{\mathbf{x}}_i(t)=\sum_{j=1}^n W_{ij}(t)\mathbf{x}_j(t)+\mathbf{e}_i(t)$ with $\mathbf{e}_i(t)=\sum_{j=1}^n W_{ij}(t)\mathbf{z}_{i,j}(t)$. In addition, $\mathbb{E}[\mathbf{e}_i(t)]=0$ and $\mathbb{E}[\|\mathbf{e}_i(t)\|^2]=\sigma^2\sum_{j=1}^n W_{ij}(t)^2\leq\sigma^2$. Hence, the conditions of Assumption 1 are satisfied.

Example 2: In many applications, there are band-limited links between the agents where the information to be sent needs to be quantized (to a certain number of bits), before transmission. The difference between the actual state and its quantized version can be modelled as the estimation noise. A popular choice for such quantizers is the stochastic quantizer [26]. In this case, it can be shown that we have $\hat{\mathbf{x}}_i(t)=\sum_{j=1}^n W_{ij}(t)\mathbf{x}_j(t)+\mathbf{e}_i(t)$, where $\mathbf{e}_i(t)$ satisfies

$$\mathbb{E}\left[\|\mathbf{e}_i(t)\|^2|\mathcal{F}_t\right]\leq\min\left(\frac{\sqrt{d}}{s},\frac{d}{s^2}\right)D.$$

Therefore, the conditions of Assumption 1 are satisfied. See the extended version of this article [25] for more detailed discussion on this.

III. Auxiliary Lemmas

In this section, we present auxiliary lemmas that play crucial roles in the proof of the main result, namely, Theorem 1 in Section IV. The proofs of Lemmas 1–5 are provided in the extended version of this article [25].

Lemma 1: Let $\{W(t)\}$ satisfy the connectivity 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 $U\in\mathbb{R}^{n\times d}$, 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 $\lambda:=\frac{\eta r_{\min}}{2Bn^2}$, $\kappa:=(1-B\lambda\beta_0)^{-1}$ and $\beta_0=\beta(1)$.

Lemma 2: For any pair of vectors $\boldsymbol{u},\mathbf{v}$, and any scalar $\theta>0$, we have $\|\boldsymbol{u}+\mathbf{v}\|^2\leq(1+\theta)\|\boldsymbol{u}\|^2+(1+\theta^{-1})\|\mathbf{v}\|^2$. Similarly, for matrices U and V and any $\theta>0$, we get

$$\|U+V\|_{\mathbf{r}}^2\leq(1+\theta)\|U\|_{\mathbf{r}}^2+\left(1+\theta^{-1}\right)\|V\|_{\mathbf{r}}.$$

Lemma 3: For any $0\leq\delta<1$ and $0<a<1$, we have

$$\prod_{k=s}^{t-1}\left(1-\frac{a}{k^\delta}\right)\leq\exp\left(-\frac{a}{1-\delta}\left(t^{1-\delta}-s^{1-\delta}\right)\right).$$

For $\delta=1$ and $0\leq a<1$, we have $\prod_{k=s}^{t-1}(1-\frac{a}{k})\leq(\frac{t}{s})^{-a}$.

Lemma 4: Let $\{\beta(t)\}$ be a sequence in $\mathbb{R}$ and λ be a nonzero scalar. Then, for all $t\geq 1$

$$\sum_{s=1}^{t-1}\beta(s)\prod_{k=s+1}^{t-1}(1-\lambda\beta(k))=\frac{1}{\lambda}-\frac{1}{\lambda}\prod_{k=1}^{t-1}(1-\lambda\beta(k)) \tag{7}$$

As a result, for any sequence $\{\beta(t)\}$ in [0,1] and $\lambda > 0$

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

Lemma 5: For any $0 \leq \delta < \min(1,\sigma)$, $0 < a \leq 1$, and every $t > \tau := (\frac{2(\sigma-\delta)}{a})^{\frac{1}{1-\delta}}$, we have

$$\sum_{s=1}^{t-1} \left[\frac{1}{s^\sigma} \prod_{k=s+1}^{t-1} \left(1 - \frac{a}{k^\delta}\right) \right] \leq A(a,\sigma,\delta) t^{-(\sigma-\delta)}$$

where $A(a,\sigma,\delta)$ is given by

$$\begin{cases} 2^\sigma \max\left\{1+\frac{2}{a}, 1+\frac{1}{\sigma-1}\left(\frac{2(\sigma-\delta)}{a}\right)^{\frac{\sigma-\delta}{1-\delta}}\right\} & \text{if } \sigma > 1 \\ 2^\sigma \max\left\{1+\frac{2}{a}, 1+\frac{2}{a}\ln\left(\frac{2(1-\delta)}{a}\right)\right\} & \text{if } \sigma = 1 \\ 2^\sigma \max\left\{1+\frac{2}{a}, 1+\frac{2(\sigma-\delta)}{a(1-\sigma)}\right\} & \text{if } 0 < \sigma < 1. \end{cases}$$

Moreover, for $\delta = 1$ and $a - \sigma + 1 \neq 0$, we have

$$\sum_{s=1}^{t-1} \left[\frac{1}{s^\sigma} \prod_{k=s+1}^{t-1} \left(1 - \frac{a}{k}\right) \right] \leq A(a,\sigma,1) t^{-\min(\sigma-1,a)}$$

where $A(a,\sigma,1) = 2^\sigma \left(1 + \frac{1}{|a-\sigma+1|}\right)$.

IV. Proof of Theorem 1

In this section, we provide the proof of the main result, namely, Theorem 1. The proof is based on the auxiliary lemmas in Section III. We prove the result in two steps. First, we bound the deviation of the agents' states from their average, and then, we analyze the distance of the average state from the global optimal point.

A. State Deviation From the Average State

The dynamics (3) is a linear time-varying system

$$X(t+1) = A(t)X(t) + U(t) \tag{8}$$

with $A(t) = (1-\beta(t))I + \beta(t)W(t)$ and the control input $U(t) = \beta(t)E(t) - \alpha(t)\beta(t)\nabla f(X(t))$. Therefore,

$$X(t) = \sum_{s=1}^{t-1} \Phi(t:s)U(s) + \Phi(t:0)X(1) \tag{9}$$

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 (8). We also define $P(t:s) := \beta(s)(\Phi(t:s) - \mathbf{1}\mathbf{r}^T)$ for the notational simplicity. As a result of Lemma 1, we have $\|P(t:s)U\|_{\mathbf{r}} \leq \pi(t:s)\|U\|_{\mathbf{r}}$, where $\pi(t:s)$ is defined by

$$\pi(t:s) := \beta(s)\kappa^{\frac{1}{2}} \prod_{k=s+1}^{t-1} (1-\lambda\beta(k))^{\frac{1}{2}}. \tag{10}$$

Assuming $X(1) = \mathbf{0}$, the dynamic in (9) reduces to

$$X(t) = \sum_{s=1}^{t-1} \Phi(t:s)U(s). \tag{11}$$

Moreover, multiplying both sides of (11) from the left by $\mathbf{r}^T$ and using the fact that $\mathbf{r}^T A(t) = \mathbf{r}^T$, we get

$$\bar{\mathbf{x}}(t) := \mathbf{r}^T X(t) = \sum_{s=1}^{t-1} \mathbf{r}^T \Phi(t:s)U(s) = \sum_{s=1}^{t-1} \mathbf{r}^T U(s). \tag{12}$$

Then, subtracting (12) from (11), and plugging the definition of $U(s)$, we have

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

Using Lemma 2 with $\theta = 1$, we get

$$\begin{aligned} &\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2 \\ &\leq 2\left\|\sum_{s=1}^{t-1} P(t:s)E(s)\right\|_{\mathbf{r}}^2 + 2\left\|\sum_{s=1}^{t-1} \alpha(s)P(t:s)\nabla f(X(s))\right\|_{\mathbf{r}}^2 \\ &= 2\sum_{s=1}^{t-1} \|P(t:s)E(s)\|_{\mathbf{r}}^2 + 2\sum_{s\neq q} \langle P(t:s)E(s), P(t:q)E(q)\rangle \\ &\quad + 2\left\|\sum_{s=1}^{t-1} \alpha(s)P(t:s)\nabla f(X(s))\right\|_{\mathbf{r}}^2. \end{aligned} \tag{13}$$

Using facts that $E(s)$ is measurable with respect to $\mathcal{F}_q$ for $q > s$ and $\mathbb{E}[E(q)|\mathcal{F}_q] = 0$, we have

$$\begin{aligned} &\mathbb{E}\left[\langle P(t:s)E(s), P(t:q)E(q)\rangle\right] \\ &\quad = \mathbb{E}\left[\mathbb{E}\left[\langle P(t:s)E(s), P(t:q)E(q)\rangle \,|\mathcal{F}_q\right]\right] \\ &\quad = \mathbb{E}\left[\langle P(t:s)E(s), P(t:q)\mathbb{E}\left[E(q)|\mathcal{F}_q\right]\rangle\right] = 0. \end{aligned} \tag{14}$$

Using a similar argument for $q < s$ and conditioning on $\mathcal{F}_s$, we conclude that (14) holds for all $q \neq s$. Therefore, taking the expectation of both sides of (13) and noting the identity in (14), we get

$$\begin{aligned} \mathbb{E}\left[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right] &\leq 2\sum_{s=1}^{t-1} \mathbb{E}\left[\|P(t:s)E(s)\|_{\mathbf{r}}^2\right] \\ &\quad + 2\mathbb{E}\left[\left\|\sum_{s=1}^{t-1} \alpha(s)P(t:s)\nabla f(X(s))\right\|_{\mathbf{r}}^2\right]. \end{aligned} \tag{15}$$

We continue with bounding the first term in (15). From Assumption 1, we have

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

This together with Lemma 1 arrives at

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

$$= \gamma\kappa \sum_{s=1}^{t-1} \left[\frac{\beta_0^2}{s^{2\mu}} \prod_{k=s+1}^{t-1} \left(1 - \lambda \frac{\beta_0}{k^\mu}\right) \right].$$

Therefore, using Lemma 5 with parameters $(\sigma, \delta, \tau) = (2\mu, \mu, T_1)$, we arrive at

$$\sum_{s=1}^{t-1} \mathbb{E}\left[\|P(t:s)E(s)\|_{\mathbf{r}}^2\right] \leq \epsilon_1 t^{-\mu} \text{for } t \geq T_1 := \left\lceil (2\mu/\lambda\beta_0)^{\frac{1}{1-\mu}} \right\rceil. \tag{16}$$

To bound the second term in (15), using the triangle inequality for norm $\|\cdot\|_{\mathbf{r}}$, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{s=1}^{t-1} \alpha(s) P(t:s) \nabla f(X(s))\right\|_{\mathbf{r}}^2\right] \\
&\leq \mathbb{E}\left[\left(\sum_{s=1}^{t-1} \|\alpha(s) P(t:s) \nabla f(X(s))\|_{\mathbf{r}}\right)^2\right] \\
&= \sum_{1\leq s,q\leq t-1} \mathbb{E}\left[\alpha(s) \|P(t:s)\nabla f(X(s))\|_{\mathbf{r}}\right. \\
&\quad \left. \times\, \alpha(q) \|P(t:q)\nabla f(X(q))\|_{\mathbf{r}}\right] \\
&\leq \sum_{1\leq s,q\leq t-1} \mathbb{E}\left[\alpha(s)\pi(t:s) \|\nabla f(X(s))\|_{\mathbf{r}}\right. \\
&\quad \left. \times\, \alpha(q)\pi(t:q) \|\nabla f(X(q))\|_{\mathbf{r}}\right] \\
&= \sum_{1\leq s,q\leq t-1} \pi(t:s)\pi(t:q)\mathbb{E}\left[\alpha(s) \|\nabla f(X(s))\|_{\mathbf{r}}\right. \\
&\quad \left. \times\, \alpha(q) \|\nabla f(X(q))\|_{\mathbf{r}}\right]
\end{aligned} \tag{17}$$

where the last inequality follows from Lemma 1 and $\pi(t:s)$ is given by (10). Using the inequality $2ab \leq a^2 + b^2$, we can further upper-bound (17) to arrive at

$$\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{s=1}^{t-1} \alpha(s) P(t:s) \nabla f(X(s))\right\|_{\mathbf{r}}^2\right] \\
&\leq \frac{1}{2} \sum_{1\leq s,q\leq t-1} \pi(t:s)\pi(t:q)\mathbb{E}\left[\alpha^2(s) \|\nabla f(X(s))\|_{\mathbf{r}}^2\right. \\
&\quad \left. + \alpha^2(q) \|\nabla f(X(q))\|_{\mathbf{r}}^2\right] \\
&= \sum_{1\leq s,q\leq t-1} \pi(t:s)\pi(t:q)\mathbb{E}\left[\alpha^2(s) \|\nabla f(X(s))\|_{\mathbf{r}}^2\right] \\
&= \left(\sum_{q=1}^{t-1} \pi(t:q)\right)\left(\sum_{s=1}^{t-1} \alpha^2(s)\pi(t:s)\mathbb{E}\left[\|\nabla f(X(s))\|_{\mathbf{r}}^2\right]\right)
\end{aligned} \tag{18}$$

But using $\sqrt{1-x} \leq 1 - x/2$ and Lemma 4, we have

$$\sum_{q=1}^{t-1} \pi(t:q) = \sum_{q=1}^{t-1} \left[\beta(q)\kappa^{\frac{1}{2}} \prod_{k=q+1}^{t-1} (1 - \lambda\beta(k))^{\frac{1}{2}}\right] \tag{19}$$

$$\leq \sum_{q=1}^{t-1} \beta(q)\kappa^{\frac{1}{2}} \prod_{k=q+1}^{t-1} \left(1 - \frac{\lambda}{2}\beta(k)\right) \leq \frac{2}{\lambda}\kappa^{\frac{1}{2}}. \tag{20}$$

Moreover, we can write

$$\begin{aligned}
\sum_{s=1}^{t-1} \alpha^2(s)\pi(t:s)\mathbb{E}\left[\|\nabla f(X(s))\|_{\mathbf{r}}^2\right] &\overset{(a)}{\leq} K \sum_{s=1}^{t-1} \alpha^2(s)\pi(t:s) \\
&\overset{(b)}{=} K \sum_{s=1}^{t-1} \alpha^2(s) \left[\beta(s)\kappa^{\frac{1}{2}} \prod_{k=s+1}^{t-1} (1 - \lambda\beta(k))^{\frac{1}{2}}\right] \\
&\overset{(c)}{\leq} K \sum_{s=1}^{t-1} \alpha^2(s)\beta(s)\kappa^{\frac{1}{2}} \prod_{k=s+1}^{t-1} \left(1 - \frac{\lambda}{2}\beta(k)\right)
\end{aligned}$$

where (a) follows from Assumption 3(c), (b) uses the definition of $\pi(t:s)$ from (10), and the inequality in (c) follows from $\sqrt{1-x} \leq 1 - x/2$ for $x \leq 1$. Therefore, using Lemma 5 with $(\sigma, \delta, \tau) = (2\nu + \mu, \mu, T_2)$, for

$$t \geq T_2 := \left\lceil (8\nu/\lambda\beta_0)^{\frac{1}{1-\mu}} \right\rceil \tag{21}$$

we arrive at

$$\begin{aligned}
&\sum_{s=1}^{t-1} \alpha^2(s)\pi(t:s)\mathbb{E}\left[\|\nabla f(X(s))\|_{\mathbf{r}}^2\right] \\
&\leq K\alpha_0^2\beta_0\kappa^{\frac{1}{2}} \sum_{s=1}^{t-1} \frac{1}{s^{2\nu+\mu}} \prod_{k=s+1}^{t-1} \left(1 - \frac{\lambda\beta_0}{2}\frac{1}{k^\mu}\right) \\
&\leq \epsilon_2 t^{-2\nu}
\end{aligned} \tag{22}$$

where $\epsilon_2 := K\alpha_0^2\beta_0\kappa^{\frac{1}{2}} A(\lambda\beta_0/2, 2\nu + \mu, \mu)$. Plugging (22) and (19) into (18) and using (17), we conclude

$$\mathbb{E}\left[\left\|\sum_{s=1}^{t-1} \alpha(s) P(t:s) \nabla f(X(s))\right\|_{\mathbf{r}}^2\right] \leq \frac{2}{\lambda}\kappa^{\frac{1}{2}}\epsilon_2 t^{-2\nu}. \tag{23}$$

Finally, using the bounds obtained in (16) and (23) in (15), for $\epsilon_3 := 2\epsilon_1 + 4\kappa^{\frac{1}{2}}\epsilon_2/\lambda$, we get

$$\begin{aligned}
\mathbb{E}\left[\|X(t) - \mathbf{1}\bar{\mathbf{x}}(t)\|_{\mathbf{r}}^2\right] &\leq 2\epsilon_1 t^{-\mu} + \frac{4}{\lambda}\kappa^{\frac{1}{2}}\epsilon_2 t^{-2\nu} \\
&\leq \left(2\epsilon_1 + \frac{4}{\lambda}\kappa^{\frac{1}{2}}\epsilon_2\right) t^{-\min(\mu, 2\nu)} \\
&= \epsilon_3 t^{-\min(\mu, 2\nu)}.
\end{aligned} \tag{24}$$

B. Average State Distance to the Optimal Point

Now, we derive an upper bound for the average distance between the mean of the agents' states and the global optimal point, i.e., $Q(t) := \mathbb{E}[\|\bar{\mathbf{x}}(t) - \mathbf{x}^\star\|^2]$, where $\mathbf{x}^\star$ is the minimizer of the function $f(\mathbf{x})$. Recall that $\bar{\mathbf{x}}(t) = \mathbf{r}^T X(t) = \sum_{i=1}^n r_i \mathbf{x}_i(t)$ and $\mathbf{r}^T W(t) = \mathbf{r}^T$. Hence, multiplying both sides of (3) by $\mathbf{r}^T$, we get

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

We define $g(t) := \mathbf{r}^T \nabla f(X(t)) = \sum_{i=1}^n r_i \nabla f_i(\mathbf{x}_i(t))$ and $\bar{g}(t) := \nabla f(\bar{\mathbf{x}}(t)) = \sum_{i=1}^n r_i \nabla f_i(\bar{\mathbf{x}}(t))$. Hence, we can write

$$\begin{aligned}
&\mathbb{E}\left[\|\bar{\mathbf{x}}(t+1) - \mathbf{x}^\star\|^2 | \mathcal{F}_t\right] \\
&= \mathbb{E}\left[\|\bar{\mathbf{x}}(t) + \beta(t)\mathbf{r}^T E(t) - \alpha(t)\beta(t)g(t) - \mathbf{x}^\star\|^2 | \mathcal{F}_t\right] \\
&= \|\bar{\mathbf{x}}(t) - \mathbf{x}^\star - \alpha(t)\beta(t)g(t)\|^2 + \mathbb{E}\left[\|\beta(t)\mathbf{r}^T E(t)\|^2 | \mathcal{F}_t\right]
\end{aligned} \tag{25}$$

where the last equality follows from the fact that $X(t)$ is measurable with respect to $\mathcal{F}_t$ and Assumption 1, implying $\mathbb{E}[\beta(t)\mathbf{r}^T E(t) | \mathcal{F}_t] = 0$, which leads to

$$2\left\langle \bar{\mathbf{x}}(t) - \alpha(t)\beta(t)g(t) - \mathbf{x}^\star, \mathbb{E}\left[\beta(t)\mathbf{r}^T E(t) | \mathcal{F}_t\right]\right\rangle = 0.$$

Taking the expectation of both sides of (25), and using the tower rule, we get

$$\begin{aligned}
Q(t+1) &= \mathbb{E}\left[\|\bar{\mathbf{x}}(t+1) - \mathbf{x}^\star\|^2\right] \\
&= \mathbb{E}\left[\|\bar{\mathbf{x}}(t) - \mathbf{x}^\star - \alpha(t)\beta(t)g(t)\|^2\right] + \mathbb{E}\left[\|\beta(t)\mathbf{r}^T E(t)\|^2\right].
\end{aligned} \tag{26}$$

In order to bound the first term in (25), we use Lemma 2 for vectors $\boldsymbol{u} = \bar{\mathbf{x}}(t) - \mathbf{x}^\star - \alpha(t)\beta(t)\bar{g}(t)$ and $\mathbf{v} = \alpha(t)\beta(t)(\bar{g}(t) - g(t))$, and parameter $\theta = \frac{\rho L}{\rho+L}\alpha(t)\beta(t)$. Hence, we can write

$$\begin{aligned}
&\|\bar{\mathbf{x}}(t) - \mathbf{x}^\star - \alpha(t)\beta(t)g(t)\|^2 \\
&= \|\bar{\mathbf{x}}(t) - \mathbf{x}^\star - \alpha(t)\beta(t)\bar{g}(t) + \alpha(t)\beta(t)\bar{g}(t) - \alpha(t)\beta(t)g(t)\|^2 \\
&\le \left(1 + \frac{\rho L}{\rho + L}\alpha(t)\beta(t)\right)\|\bar{\mathbf{x}}(t) - \mathbf{x}^\star - \alpha(t)\beta(t)\bar{g}(t)\|^2 \\
&\quad + \alpha(t)\beta(t)\left(\alpha(t)\beta(t) + \frac{\rho + L}{\rho L}\right)\|\bar{g}(t) - g(t)\|^2. \qquad (27)
\end{aligned}$$

Next, we use [27, Th. 2.1.11] to bound the first term in (27). Note that Assumptions 3(a) and 3(b) (and Remark 1) guarantee that the conditions of [27, Th. 2.1.11] are satisfied. Thus, we have

$$\langle \bar{\mathbf{x}}(t) - \mathbf{x}^\star, \nabla f(\bar{\mathbf{x}}(t))\rangle \ge c_1\|\nabla f(\bar{\mathbf{x}}(t))\|^2 + c_2\|\bar{\mathbf{x}}(t) - \mathbf{x}^\star\|^2$$

or equivalently,

$$\langle \bar{\mathbf{x}}(t) - \mathbf{x}^\star, \bar{g}(t)\rangle \ge c_1\|\bar{g}(t)\|^2 + c_2\|\bar{\mathbf{x}}(t) - \mathbf{x}^\star\|^2 \qquad (28)$$

where $c_1 = \frac{1}{\rho+L}$ and $c_2 = \frac{\rho L}{\rho+L}$. Therefore, for the first term in (27), we can write

$$\begin{aligned}
&\|\bar{\mathbf{x}}(t) - \mathbf{x}^\star - \alpha(t)\beta(t)\bar{g}(t)\|^2 \\
&\quad = \|\bar{\mathbf{x}}(t) - \mathbf{x}^\star\|^2 + \alpha^2(t)\beta^2(t)\|\bar{g}(t)\|^2 \\
&\qquad - 2\alpha(t)\beta(t)\langle \bar{\mathbf{x}}(t) - \mathbf{x}^\star, \bar{g}(t)\rangle \\
&\quad \le (1 - 2c_2\alpha(t)\beta(t))\|\bar{\mathbf{x}}(t) - \mathbf{x}^\star\|^2 \\
&\qquad + \alpha(t)\beta(t)(\alpha(t)\beta(t) - 2c_1)\|\bar{g}(t)\|^2. \qquad (29)
\end{aligned}$$

Let us set

$$T_3 := \left\lceil \left(\frac{\alpha_0\beta_0}{2c_1}\right)^{\frac{1}{\mu+\nu}} \right\rceil = \left\lceil \left(\frac{\alpha_0\beta_0(\rho+L)}{2}\right)^{\frac{1}{\mu+\nu}} \right\rceil \qquad (30)$$

such that $\alpha(t)\beta(t) \le 2c_1$ for any $t \ge T_3$. Hence, for $t \ge T_3$, the second term in (29) is nonpositive, and thus

$$\|\bar{\mathbf{x}}(t) - \mathbf{x}^\star - \alpha(t)\beta(t)\bar{g}(t)\|^2 \le (1 - 2c_2\alpha(t)\beta(t))\|\bar{\mathbf{x}}(t) - \mathbf{x}^\star\|^2.$$

Taking expectation from both sides, we get

$$\mathbb{E}\left[\|\bar{\mathbf{x}}(t) - \mathbf{x}^\star - \alpha(t)\beta(t)\bar{g}(t)\|^2\right] \le (1 - 2c_2\alpha(t)\beta(t))Q(t). \qquad (31)$$

The average of the second term in (27) can be bounded as

$$\begin{aligned}
\mathbb{E}\left[\|\bar{g}(t) - g(t)\|^2\right] &= \mathbb{E}\left[\left\|\sum_{i=1}^n r_i\left(\left(\nabla f_i(\bar{\mathbf{x}}(t)) - \nabla f_i(\mathbf{x}_i(t))\right)\right)\right\|^2\right] \\
&\overset{(a)}{\le} \mathbb{E}\left[\sum_{i=1}^n r_i\|\nabla f_i(\bar{\mathbf{x}}(t)) - \nabla f_i(\mathbf{x}_i(t))\|^2\right] \\
&\overset{(b)}{\le} L^2\sum_{i=1}^n r_i\mathbb{E}\left[\|\bar{\mathbf{x}}(t) - \mathbf{x}_i(t)\|^2\right] \\
&= L^2\mathbb{E}\left[\|\mathbf{1}\bar{\mathbf{x}}(t) - X(t)\|_{\mathbf{r}}^2\right] \overset{(c)}{\le} L^2\epsilon_3 t^{-\min(\mu,2\nu)} \qquad (32)
\end{aligned}$$

where (a) follows from the convexity of $\|\cdot\|^2$, (b) holds due to Assumption 3(a), and we used (24) for (c).

Taking expectation from both sides of (27) and recalling that $c_2 = \rho L/(\rho + L)$, we arrive at

$$\begin{aligned}
&\mathbb{E}\left[\|\bar{\mathbf{x}}(t) - \mathbf{x}^\star - \alpha(t)\beta(t)g(t)\|^2\right] \\
&\le (1 + c_2\alpha(t)\beta(t))\,\mathbb{E}\left[\|\bar{\mathbf{x}}(t) - \mathbf{x}^\star - \alpha(t)\beta(t)\bar{g}(t)\|^2\right] \\
&\quad + \alpha(t)\beta(t)\,(\alpha(t)\beta(t) + 1/c_2)\,\mathbb{E}\left[\|\bar{g}(t) - g(t)\|^2\right] \\
&\overset{(d)}{\le} (1 + c_2\alpha(t)\beta(t))(1 - 2c_2\alpha(t)\beta(t)Q(t) \\
&\quad + \alpha(t)\beta(t)(\alpha(t)\beta(t) + 1/c_2)L^2\epsilon_3 t^{-\min(\mu,2\nu)} \\
&\overset{(e)}{\le} (1 - c_2\alpha(t)\beta(t))Q(t) \\
&\quad + \alpha(t)\beta(t)(\alpha(t)\beta(t) + 1/c_2)L^2\epsilon_3 t^{-\min(\mu,2\nu)} \qquad (33)
\end{aligned}$$

where the inequality in (d) follows from (31) and (32), and (e) holds since

$$(1 + c_2\alpha(t)\beta(t))(1 - 2c_2\alpha(t)\beta(t)) \le 1 - c_2\alpha(t)\beta(t).$$

From Assumption 1, we get

$$\begin{aligned}
\left[\mathbb{E}\left[E(t)E(t)^T|\mathcal{F}_t\right]\right]_{ij} &= \mathbb{E}\left[\mathbf{e}_i(t)\mathbf{e}_j^T(t)|\mathcal{F}_t\right] \\
&\le \sqrt{\mathbb{E}\left[\|\mathbf{e}_i(t)\|^2|\mathcal{F}_t\right]\mathbb{E}\left[\|\mathbf{e}_j(t)\|^2|\mathcal{F}_t\right]} \le \gamma \qquad (34)
\end{aligned}$$

for all $1 \le i, j \le n$. Thus, for the second term in (25), we arrive at

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{r}^T E(t)\right\|^2|\mathcal{F}_t\right] &= \mathbf{r}^T\mathbb{E}\left[E(t)E(t)^T|\mathcal{F}_t\right]\mathbf{r} \\
&\le \mathbf{r}^T\left(\gamma\mathbf{1}\mathbf{1}^T\right)\mathbf{r} = \gamma \qquad (35)
\end{aligned}$$

as $\mathbf{r}^T\mathbf{1} = 1$. Taking expectations from both sides of (35), and using the tower rule, we arrive at

$$\mathbb{E}\left[\left\|\mathbf{r}^T E(t)\right\|^2\right] = \mathbb{E}\left[\mathbb{E}\left[\left\|\mathbf{r}^T E(t)\right\|^2|\mathcal{F}_t\right]\right] \le \gamma. \qquad (36)$$

Using (33) and (36) in (26), we can write

$$\begin{aligned}
Q(t+1) &\le (1 - c_2\alpha(t)\beta(t))Q(t) \\
&\quad + \alpha(t)\beta(t)(\alpha(t)\beta(t) + 1/c_2)L^2\epsilon_3 t^{-\min(\mu,2\nu)} \\
&\quad + \gamma\beta^2(t) \\
&\le (1 - c_2\alpha(t)\beta(t))Q(t) + \epsilon_4 t^{-\min(2\mu,3\nu+\mu)} \qquad (37)
\end{aligned}$$

where $\epsilon_4 := \alpha_0\beta_0(\alpha_0\beta_0 + 1/c_2)L^2\epsilon_3 + \gamma\beta_0^2$ and the last inequality follows from

$$\begin{aligned}
&\alpha(t)\beta(t)(\alpha(t)\beta(t) + 1/c_2)L^2\epsilon_3 t^{-\min(\mu,2\nu)} + \gamma\beta^2(t) \\
&\le \alpha_0\beta_0 t^{-\nu-\mu}(\alpha_0\beta_0 + 1/c_2)L^2\epsilon_3 t^{-\min(\mu,2\nu)} + \gamma\beta_0^2 t^{-2\mu} \\
&= \alpha_0\beta_0(\alpha_0\beta_0 + 1/c_2)L^2\epsilon_3 t^{-\min(2\mu+\nu,3\nu+\mu)} + \gamma\beta_0^2 t^{-2\mu} \\
&\le (\alpha_0\beta_0(\alpha_0\beta_0 + 1/c_2)L^2\epsilon_3 + \gamma\beta_0^2)t^{-\min(2\mu,3\nu+\mu)}.
\end{aligned}$$

Consider a general dynamic $Q(t)$ satisfying a recursive inequality $Q(t+1) \le G(t)Q(t) + H(t)$ for every $t \ge T_0 := \max(T_1, T_2, T_3)$, in which $G(k) \ge 0$ for all k. Then, we have

$$Q(T) \le \left(\prod_{k=T_0}^{T-1} G(k)\right)Q(T_0) + \sum_{s=T_0}^{T-1} H(s)\left[\prod_{k=s+1}^{T-1} G(k)\right].$$

for any starting time T_0. Therefore, (37) yields

$$Q(T) \le \prod_{k=T_0}^{T-1}(1 - c_2\alpha(k)\beta(k))Q(T_0)$$

$$+\epsilon_4\sum_{s=T_0}^{T-1}s^{-\min(2\mu,3\nu+\mu)}\left[\prod_{k=s+1}^{T-1}(1-c_2\alpha(k)\beta(k))\right]$$
$$=\prod_{k=T_0}^{T-1}\left(1-\frac{c_2\alpha_0\beta_0}{k^{\nu+\mu}}\right)Q(T_0)$$
$$+\epsilon_4\sum_{s=T_0}^{T-1}s^{-\min(2\mu,3\nu+\mu)}\left[\prod_{k=s+1}^{T-1}\left(1-\frac{c_2\alpha_0\beta_0}{k^{\nu+\mu}}\right)\right]$$
$$\le\prod_{k=T_0}^{T-1}\left(1-\frac{c_2\alpha_0\beta_0}{k^{\nu+\mu}}\right)Q(T_0)$$
$$+\epsilon_4\sum_{s=1}^{T-1}s^{-\min(2\mu,3\nu+\mu)}\left[\prod_{k=s+1}^{T-1}\left(1-\frac{c_2\alpha_0\beta_0}{k^{\nu+\mu}}\right)\right]. \quad (38)$$

To further upper-bound the right-hand side of (38), we may distinguish two individual cases. When $\mu+\nu<1$, we can define

$$T_4:=\left\lceil\left(\frac{2\min(\mu-\nu,2\nu)}{c_2\alpha_0\beta_0}\right)^{\frac{1}{1-\mu-\nu}}\right\rceil. \quad (39)$$

Then, we can apply Lemmas 3 with $\delta=\mu+\nu$ on the first term of (38), and use Lemma 5 with $(\sigma,\delta,\tau)=(\min(2\mu,3\nu+\mu),\nu+\mu,T_4)$ for the second term of (38). As a result, for any $T\ge\max(T_0,T_4)$, we have

$$Q(T)\le\exp\left(-\frac{c_2\alpha_0\beta_0}{1-\mu-\nu}\left(T^{1-\mu-\nu}-T_3^{1-\mu-\nu}\right)\right)Q(T_0)$$
$$+\epsilon_4\epsilon_5T^{-\min(\mu-\nu,2\nu)} \quad (40)$$

where $\epsilon_5:=A(c_2\alpha_0\beta_0,\min(2\mu,3\nu+\mu),\nu+\mu)$.

Next, when $\mu+\nu=1$, similar to the previous case, we use Lemma 3 to upper-bound the first term of (38), and apply Lemma 5 with $(\sigma,\delta,\tau)=(\min(2\mu,3\nu+\mu),1,T_4)$ on its second term. This leads to

$$Q(T)\le\prod_{k=T_0}^{T-1}\left(1-\frac{c_2\alpha_0\beta_0}{k}\right)Q(T_0)$$
$$+\epsilon_4\sum_{s=1}^{T-1}s^{-\min(2\mu,3\nu+\mu)}\left[\prod_{k=s+1}^{T-1}\left(1-\frac{c_2\alpha_0\beta_0}{k}\right)\right]$$
$$\le\left(\frac{T}{T_0}\right)^{-c_2\alpha_0\beta_0}Q(T_0)+\epsilon_4\epsilon_5T^{-\min(2\mu-1,3\nu+\mu-1,c_2\alpha_0\beta_0)}$$
$$=\left(\frac{T}{T_0}\right)^{-c_2\alpha_0\beta_0}Q(T_0)+\epsilon_4\epsilon_5T^{-\min(\mu-\nu,2\nu,c_2\alpha_0\beta_0)}$$
$$\le\left(T_0^{c_2\alpha_0\beta_0}Q(T_0)+\epsilon_4\epsilon_5\right)T^{-\min(\mu-\nu,2\nu)} \quad (41)$$

where the last inequality holds as $c_2\alpha_0\beta_0\ge\min(\mu-\nu,2\nu)$.

C. Total State Deviation From the Optimum Solution

Combining the above bounds, we can conclude the proof of Theorem 1. In particular, for $\mu+\nu<1$, we have

$$\mathbb{E}\left[\|X(T)-\mathbf{1}\mathbf{x}^\star\|_{\mathbf{r}}^2\right]$$
$$=\mathbb{E}\left[\|X(T)-\mathbf{1}\bar{\mathbf{x}}(T)+\mathbf{1}\bar{\mathbf{x}}(T)-\mathbf{1}\mathbf{x}^\star\|_{\mathbf{r}}^2\right]$$
$$\le2\left(\mathbb{E}\left[\|X(T)-\mathbf{1}\bar{\mathbf{x}}(T)\|_{\mathbf{r}}^2\right]+\mathbb{E}\left[\|\mathbf{1}\bar{\mathbf{x}}(T)-\mathbf{1}\mathbf{x}^\star\|_{\mathbf{r}}^2\right]\right)$$
$$\le2\epsilon_3T^{-\min(\mu,2\nu)}$$
$$+2\exp\left(-\frac{c_2\alpha_0\beta_0}{1-\mu-\nu}\left(T^{1-\mu-\nu}-T_0^{1-\mu-\nu}\right)\right)Q(T_0)$$
$$+2\epsilon_4\epsilon_5T^{-\min(\mu-\nu,2\nu)}.$$

for every $T\ge\max(T_0,T_4)=\max(T_1,T_2,T_3,T_4)$. This implies Theorem 1 for $\mu+\nu<1$. Note that

$$\xi_1:=4\gamma\kappa\beta_0^2A(\lambda\beta_0,2\mu,\mu)+\frac{8K\kappa\alpha_0^2\beta_0}{\lambda}A(\lambda\beta_0/2,2\nu+\mu,\mu)$$
$$\xi_2:=2\exp\left(\frac{\alpha_0\beta_0\rho L}{(1-\mu-\nu)(\rho+L)}T_0^{1-\mu-\nu}\right)Q(T_0)$$
$$\xi_3:=\frac{\alpha_0\beta_0\rho L}{(1-\mu-\nu)(\rho+L)}$$
$$\xi_4:=\left(\alpha_0\beta_0(\alpha_0\beta_0\rho L+\rho+L)\xi_1L/\rho+2\gamma\beta_0^2\right)$$
$$\times A\left(\frac{\alpha_0\beta_0\rho L}{\rho+L},\min(2\mu,3\nu+\mu),\mu+\nu\right). \quad (42)$$

and $A(\cdot,\cdot,\cdot)$ is defined in Lemma 5. Similarly, for $\mu+\nu=1$,

$$\mathbb{E}\left[\|X(T)-\mathbf{1}\mathbf{x}^\star\|_{\mathbf{r}}^2\right]$$
$$\le2\epsilon_3T^{-\min(\mu,2\nu)}+2\left(T_0^{c_2\alpha_0\beta_0}Q(T_0)+\epsilon_4\epsilon_5\right)T^{-\min(\mu-\nu,2\nu)}.$$

for every $T\ge T_0=\max(T_1,T_2,T_3)$. This leads to (6), where

$$\xi_5:=2T_0^{\alpha_0\beta_0\rho L/(\rho+L)}Q(T_0)+\xi_4. \quad (43)$$

V. Conclusion

We have studied distributed optimization over time-varying networks suffering from noisy and imperfect sharing of information. Inspired by the original averaging-based distributed optimization algorithm with the diminishing step-size, we showed that for the class of strongly convex cost functions, including a damping mechanism for the imperfect incoming information from neighboring agents leads to convergence to the optimizer in L_2 sense for various choices of the damping and diminishing step-size parameters. In addition, we obtained a convergence rate as a function of these parameters. Optimizing the resulting rate over the set of feasible parameters leads to the convergence rate $\mathcal{O}(T^{-1/2})$.

References

[1] M. Rabbat and R. Nowak, "Distributed optimization in sensor networks," in *Proc. 3rd Int. Symp. Inf. Process. Sensor Netw.*, 2004, pp. 20–27.

[2] S. Kar and J. M. Moura, "Distributed consensus algorithms in sensor networks with imperfect communication: Link failures and channel noise," *IEEE Trans. Signal Process.*, vol. 57, no. 1, pp. 355–369, Jan. 2009.

[3] G. Neglia, G. Reina, and S. Alouf, "Distributed gradient optimization for epidemic routing: A preliminary evaluation," in *Proc. 2nd IFIP Wirel. Days*, 2009, pp. 1–6.

[4] K. Tsianos, S. Lawlor, and M. Rabbat, "Consensus-based distributed optimization: Practical issues and applications in large-scale machine learning," in *Proc. Annu. Allerton Conf. Commun. Control Comput.*, 2012, pp. 1543–1550.

[5] S. S. Ram, V. V. Veeravalli, and A. Nedic, "Distributed non-autonomous power control through distributed convex optimization," in *Proc. IEEE Int. Conf. Comput. Commun.*, 2009, pp. 3001–3005.

[6] L. Xiao and S. Boyd, "Optimal scaling of a gradient method for distributed resource allocation," *J. Optim. Theory Appl.*, vol. 129, no. 3, pp. 469–488, 2006.

[7] A. Ribeiro, "Ergodic stochastic optimization algorithms for wireless communication and networking," *IEEE Trans. Signal Process.*, vol. 58, no. 12, pp. 6369–6386, Dec. 2010.

[8] I. D. Schizas, A. Ribeiro, and G. B. Giannakis, "Consensus in ad hoc WSNs with noisy links—Part I: Distributed estimation of deterministic signals," *IEEE Trans. Signal Process.*, vol. 56, no. 1, pp. 350–364, Jan. 2008.
[9] A. Nedic, A. Olshevsky, A. Ozdaglar, and J. N. Tsitsiklis, "On distributed averaging algorithms and quantization effects," *IEEE Trans. Autom. Control*, vol. 54, no. 11, pp. 2506–2517, Nov. 2009.
[10] A. Nedic, A. Ozdaglar, and P. A. Parrilo, "Constrained consensus and optimization in multi-agent networks," *IEEE Trans. Autom. Control*, vol. 55, no. 4, pp. 922–938, Apr. 2010.
[11] S. Boyd, N. Parikh, and E. Chu, *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*. Boston, MA, USA: Now Publ., 2011.
[12] D. Jakovetić, J. Xavier, and J. M. Moura, "Fast distributed gradient methods," *IEEE Trans. Autom. Control*, vol. 59, no. 5, pp. 1131–1146, May 2014.
[13] K. Yuan, Q. Ling, and W. Yin, "On the convergence of decentralized gradient descent," *SIAM J. Optim.*, vol. 26, no. 3, pp. 1835–1854, 2016.
[14] J. C. Duchi, A. Agarwal, and M. J. Wainwright, "Dual averaging for distributed optimization: Convergence analysis and network scaling," *IEEE Trans. Autom. Control*, vol. 57, no. 3, pp. 592–606, Mar. 2012.
[15] A. Aghajan and B. Touri, "Distributed optimization over dependent random networks," 2020, *arXiv:2010.01956*.
[16] G. Qu and N. Li, "Harnessing smoothness to accelerate distributed optimization," *IEEE Trans. Control. Netw. Syst.*, vol. 5, no. 3, pp. 1245–1260, Sep. 2018.
[17] C. Xi and U. A. Khan, "DEXTRA: A fast algorithm for optimization over directed graphs," *IEEE Trans. Autom. Control*, vol. 62, no. 10, pp. 4980–4993, Oct. 2017.
[18] T. Tatarenko and B. Touri, "Non-convex distributed optimization," *IEEE Trans. Autom. Control*, vol. 62, no. 8, pp. 3744–3757, Aug. 2017.
[19] J. Zeng and W. Yin, "On nonconvex decentralized gradient descent," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2834–2848, Jun. 2018.
[20] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Trans. Autom. Control*, vol. 54, no. 1, pp. 48–61, Jan. 2009.
[21] A. Nedić and A. Olshevsky, "Distributed optimization over time-varying directed graphs," *IEEE Trans. Autom. Control*, vol. 60, no. 3, pp. 601–615, Mar. 2015.
[22] A. Reisizadeh, A. Mokhtari, H. Hassani, and R. Pedarsani, "An exact quantized decentralized gradient descent algorithm," *IEEE Trans. Signal Process.*, vol. 67, no. 19, pp. 4934–4947, Oct. 2019.
[23] M. Vasconcelos, T. Doan, and U. Mitra, "Improved convergence rate for a distributed two-time-scale gradient method under random quantization," 2021, *arXiv:2105.14089*.
[24] K. Srivastava and A. Nedic, "Distributed asynchronous constrained stochastic optimization," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 4, pp. 772–790, Aug. 2011.
[25] H. Reisizadeh, B. Touri, and S. Mohajer, "Distributed optimization over time-varying graphs with imperfect sharing of information: Extended version," 2022, *arXiv:2106.08469*.
[26] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "QSGD: Communication-efficient SGD via gradient quantization and encoding," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 1709–1720.
[27] Y. Nesterov, "Introductory lectures on convex programming volume I: Basic course," *Lecture Notes*, vol. 3, no. 4, 1998, Art. no. 5.