Gradient-tracking based Distributed Optimization with Guaranteed Optimality under Noisy Information Sharing

Yongqiang Wang, Tamer Başar

***Abstract*—Distributed optimization enables networked agents to cooperatively solve a global optimization problem even with each participating agent only having access to a local partial view of the objective function. Despite making significant inroads, most existing results on distributed optimization rely on noise-free information sharing among the agents, which is problematic when communication channels are noisy, messages are coarsely quantized, or shared information are obscured by additive noise for the purpose of achieving differential privacy. The problem of information-sharing noise is particularly pronounced in the state-of-the-art gradient-tracking based distributed optimization algorithms, in that information-sharing noise will accumulate with iterations on the gradient-tracking estimate of these algorithms, and the ensuing variance will even grow unbounded when the noise is persistent. This paper proposes a new gradient-tracking based distributed optimization approach that can avoid information-sharing noise from accumulating in the gradient estimation. The approach is applicable even when the inter-agent interaction is time-varying, which is key to enable the incorporation of a decaying factor in inter-agent interaction to gradually eliminate the influence of information-sharing noise. In fact, we rigorously prove that the proposed approach can ensure the almost sure convergence of all agents to the same optimal solution even in the presence of persistent information-sharing noise. The approach is applicable to general directed graphs. It is also capable of ensuring the almost sure convergence of all agents to an optimal solution when the gradients are noisy, which is common in machine learning applications. Numerical simulations confirm the effectiveness of the proposed approach.**

I. Introduction

We consider a distributed convex optimization problem where multiple agents cooperatively solve a global optimization problem via local computations and local sharing of information. This is motivated by the problem's broad applications in cooperative control [1], distributed sensing [2], multi-agent systems [3], sensor networks [4], and large-scale machine learning [5]. In many of these applications, each agent has access to only a local portion of the objective function. Such a distributed optimization problem can be formulated in the following general form:

$$\min_{\theta\in\mathbb{R}^d} F(\theta) \triangleq \frac{1}{m}\sum_{i=1}^{m} f_i(\theta) \tag{1}$$

The work of the first author was supported in part by the National Science Foundation under Grants ECCS-1912702, CCF-2106293, CCF-2215088, and CNS-2219487. Research of the second author was supported in part by the ONR MURI Grant N00014-16-1-2710 and in part by the Army Research Laboratory, under Cooperative Agreement Number W911NF-17-2-0196.

Yongqiang Wang is with the Department of Electrical and Computer Engineering, Clemson University, Clemson, SC 29634, USA yongqiw@clemson.edu

Tamer Başar is with the Coordinated Science Lab, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA basar1@illinois.edu

where m is the number of agents, $\theta \in \mathbb{R}^d$ is a decision variable common to all agents, while $f_i : \mathbb{R}^d \to \mathbb{R}$ is a local objective function private to agent i.

To solve problem (1) in a distributed manner, plenty of algorithms have been reported since the seminal works in the 1980s [6]. Some of the popular algorithms include gradient methods (e.g., [7], [8], [9], [10], [11]), distributed alternating direction method of multipliers (e.g., [12], [13]), and distributed Newton methods (e.g., [14], [15]). In this paper, we focus on distributed gradient methods, which are particulary appealing for agents with limited computational or storage capabilities due to their low computation complexity and storage requirement. Existing distributed gradient methods can be generally divided into two categories. The first category of distributed gradient methods directly concatenate gradient based steps with a consensus operation of the optimization variable (referred to as the static-consensus based approach hereafter), with typical examples including [7], [16]. These approaches are simple and efficient in computation since they require each agent to share only one variable in each iteration. However, they are only applicable in undirected graphs and directed graphs that are balanced (the sum of each agent's in-neighbor coupling weights equal to the sum of its out-neighbor coupling weights). The second category of distributed gradient methods exploit a dynamic-consensus mechanism to track the global gradient (and hence usually called gradient-tracking based approach), and are applicable to general directed graphs (see, e.g., [9], [10], [11], [17], [18]). Such approaches can ensure convergence to an optimal solution under constant stepsizes and, hence, can achieve faster convergence. However, these approaches need every agent to maintain and share an additional gradient-tracking variable besides the optimization variable, which doubles the communication overhead compared with the approaches in the first category.

Although plenty of inroads have been made in distributed optimization, most of the existing approaches assume noise-free information sharing, i.e., every agent is able to acquire neighboring agents' intermediate local optimization variables accurately without any distortion or corruption. Such an assumption does not hold any longer, however, in many application scenarios. For example, when communication channels are noisy, every message will be distorted by channel noise which is usually modeled as additive Gaussian noise [19], [20]. An even more pervasive source of noise in information sharing comes from quantization in digital communication, which maps continuous-amplitude (analog) signals into discrete-amplitude (digital) signals, and hence leads to rounding errors (so called quantization errors) on shared messages. Such quantization errors are usually modeled as additive noises and

are non-negligible when the quantizer has a limited number of quantization levels [21]. In fact, in deep learning applications where the dimension of optimization variables can scale to hundreds of millions [22], many distributed optimization algorithms purposely employ a coarse quantizer to reduce the communication overhead [23], [24], [25], resulting in large quantization errors. Furthermore, as privacy becomes an increasingly pressing need while conventional distributed optimization algorithms are proven to leak information of participating agents [13], [26], [27], [28], many privacy-aware distributed optimization algorithms opt to inject additive noise in shared messages to ensure differential privacy [29], [30], [31], [32]. The differential-privacy induced additive noise is persistent throughout the optimization process to ensure a strong privacy protection and will significantly reduce the optimization accuracy of existing distributed optimization algorithms.

In recent years, adding a decaying factor on the coupling weight has been proven effective in suppressing the influence of persistent information-sharing noise in distributed optimization [20], [33], [34], [35], [36], [37]. In combination with a decaying stepsize in gradient descent (to alleviate the effect of information-sharing noise on gradient directions), these approaches can achieve almost sure convergence to an optimal solution. However, these results are only applicable to static-consensus based distributed optimization algorithms, which work on symmetric or balanced graphs but cannot be applied to general directed interaction graphs. In fact, in gradient-tracking based distributed optimization algorithms, information-sharing noise will accumulate on the estimate of the global gradient, and its variance will grow to infinity when the information-sharing noise is persistent, as will be explained later in Sec. III. Recently, [38] proposed an algorithm which can avoid information-sharing noise from accumulating on the global-gradient estimate when the inter-agent interaction is constant. However, when the inter-agent interaction is time-varying, the approach cannot avoid noise accumulation from happening, which precludes the possibility to incorporate a decaying factor to attenuate the influence of noise. Directly combining conventional gradient-tracking based approaches with a decaying factor can reduce the speed of such noise accumulation but is unable to avoid the noise from accumulating on the estimate of the global gradient and the gradient-estimation noise variance from escaping to infinity. Our recent result exploited heterogeneous decaying factors for the optimization variable and the gradient-tracking variable, and managed to avoid the accumulated gradient-estimation noise variance from growing to infinity [39]. However, the gradient-tracking noise still accumulates with iterations, which significantly affects the accuracy of distributed optimization. In this paper, we propose to revise the mechanics of the gradient-tracking based approach to tackle information-sharing noise in distributed optimization. More specifically, we propose a new gradient-tracking based architecture which can avoid noise from accumulation on every agent's estimate of the global gradient. This approach is applicable even when the inter-agent interaction is time-varying, which enables the incorporation of a decaying factor to attenuate the influence of noise. In fact, by choosing the decaying factor appropriately, the proposed approach can gradually eliminate the influence of information-sharing noise on all agents' local gradient-estimates even when the noise is persistent, and hence ensures the final optimality of distributed optimization.

The main contributions of this paper are as follows: 1) We propose a new gradient-tracking based distributed optimization architecture that can avoid the accumulation of information-sharing noise in the estimate of the global gradient. Different from [38], which is only applicable when the inter-agent interaction is time-invariant, the new architecture allows inter-agent interaction to be time-varying, which is key to enable the incorporation of a decaying factor in the interaction to gradually eliminate the influence of persistent information-sharing noise; 2) By incorporating a decaying factor in the inter-agent interaction, we arrive at two new algorithms that are able to gradually eliminate the influence of information-sharing noise on both consensus and global-gradient estimation, which, to our knowledge, has not been achieved before. The first algorithm requires each agent to have access to the left eigenvector of the coupling weight matrix, whereas the second algorithm uses a local eigenvector estimator to avoid requiring such global information; 3) We prove that even under persistent information-sharing noise, the proposed algorithms can guarantee every agent's almost sure convergence to an optimal solution on general directed graphs. This is in contrast to existing static-consensus based algorithms in [20], [35], [36] that are only applicable to balanced directed graphs (the sum of each agent's in-neighbor coupling weights equal to the sum of its out-neighbor coupling weights); 4) We prove that the proposed approach can ensure all agents' almost sure convergence to an optimal solution even when the gradient is subject to noise, which is a common problem in machine learning applications; 5) The proposed convergence analysis has fundamental differences from existing proof techniques for gradient-tracking based algorithms. More specifically, existing convergence analysis of gradient-tracking based algorithms relies on formulating the error dynamics as a linear time-invariant system of inequalities, whose convergence is determined by a constant systems matrix (even under time-varying coupling graphs [40]). For example, in the existing gradient-tracking based distributed optimization algorithms, this constant systems matrix is denoted as A in [18], [41], J in [11], G in [42], or M in [40]. Then, existing analysis establishes exponential (linear) convergence by proving that the spectral radius of this systems matrix is a constant value strictly less than one. However, under a decaying coupling strength, the spectral radius of the systems matrix in the conventional formulation will converge to one, which makes it impossible to use the conventional spectral-radius based analysis. Therefore, to prove convergence of our algorithms, we propose a new martingale convergence theorem based approach, which is fundamentally different from conventional proof techniques for gradient-tracking based optimization algorithms. Moreover, our algorithms and theoretical derivations only require the objective functions to be convex and Lipschitz continuous in gradients, which is different from many existing results that require the objective functions to be coercive [9] or strongly

convex [38], [42], or to have bounded gradients [20], [24], [35], [36], [43].

The rest of the paper is organized as follows. Sec. II formulates the problem and provides some results for a later use. Sec. III presents a dynamic-consensus based gradient-tracking method that can avoid noise accumulation on the gradient-tracking estimate. Sec. IV establishes the almost sure convergence of all agents to a same optimal solution. Sec. V extends the results by incorporating a left-eigenvector estimator into each agent's local update, which ensures a decentralized implementation of the approach even when information of the coupling weight matrices is not locally available to individual agents. Sec. VI extends the results to the case where the gradient is subject to noise and establishes almost sure convergence of all agents to an optimal solution. Sec. VII presents numerical comparisons with existing gradient methods to corroborate the theoretical results. Finally, Sec. VIII concludes the paper.

Notations: We use $\mathbb{R}^d$ to denote the Euclidean space of dimension d. We write I_d for the identity matrix of dimension d, and $\mathbf{1}_d$ for the d-dimensional column vector will all entries equal to 1; in both cases we suppress the dimension when clear from the context. A vector is viewed as a column vector. For a vector x, x_i denotes its ith element. We use $\langle\cdot,\cdot\rangle$ to denote the inner product of two vectors and $\|x\|$ for the standard Euclidean norm of a vector x. We write $\|A\|$ for the matrix norm induced by the vector norm $\|\cdot\|$, unless stated otherwise. We let A^T denote the transpose of a matrix A. We also use other vector/matrix norms defined under a certain transformation determined by a matrix W, which will be represented as $\|\cdot\|_W$. A matrix is column-stochastic when its entries are nonnegative and elements in every column add up to one. A matrix A is said to be row-stochastic when its entries are nonnegative and elements in every row add up to one. For two vectors u and v with the same dimension, we use $u \leq v$ to represent that every entry of u is no larger than the corresponding entry of v. Often, we abbreviate *almost surely* by *a.s.*

II. Problem Formulation and Preliminaries

We consider a network of m agents. The agents interact on a general directed graph. We describe a directed graph using an ordered pair $\mathcal{G} = ([m], \mathcal{E})$, where $[m] = \{1, 2, \ldots, m\}$ is the set of nodes (agents) and $\mathcal{E} \subseteq [m] \times [m]$ is the edge set of ordered node pairs describing the interaction among agents. For a nonnegative weight matrix $W = \{w_{ij}\} \in \mathbb{R}^{m\times m}$, we define the induced directed graph as $\mathcal{G}_W = ([m], \mathcal{E}_W)$, where the directed edge (i, j) from agent j to agent i exists, i.e., $(i, j) \in \mathcal{E}_W$ if and only if $w_{ij} > 0$. For an agent $i \in [m]$, its in-neighbor set $\mathbb{N}_i^{\text{in}}$ is defined as the collection of agents j such that $w_{ij} > 0$; similarly, the out-neighbor set $\mathbb{N}_i^{\text{out}}$ of agent i is the collection of agents j such that $w_{ji} > 0$.

By assigning a copy x_i of the decision variable x to each agent i, and then imposing the requirement $x_i = x$ for all $1 \leq i \leq m$, we can rewrite the optimization problem (1) as the following equivalent multi-agent optimization problem:

$$\min_{x\in\mathbb{R}^{md}} f(x) \triangleq \frac{1}{m}\sum_{i=1}^{m} f_i(x_i) \text{ s.t. } x_1 = x_2 = \cdots = x_m \quad (2)$$

where $x_i \in \mathbb{R}^d$ is agent i's decision variable and the collection of the agents' variables is

$$x = [x_1^T, x_2^T, \ldots, x_m^T]^T \in \mathbb{R}^{md}.$$

We make the following standard assumption on the individual objective functions:

Assumption 1. *Problem (1) has at least one optimal solution θ^*. Every $f_i(\cdot)$ is convex and has Lipschitz continuous gradients, i.e., for some $L > 0$, we have*

$$\|\nabla f_i(u) - \nabla f_i(v)\| \leq L\|u - v\|, \quad \forall i \text{ and } \forall u, v \in \mathbb{R}^d.$$

III. The Proposed Approach

In gradient-tracking based algorithms, besides an optimization variable x_i^k, every agent $i \in [m]$ also maintains and updates a gradient-tracking variable y_i^k which estimates the global gradient ("joint agent" descent direction). Both the optimization variable and the gradient-tracking variable have to be shared with neighboring agents. The two variables can be shared using two different communication networks, usually called, $\mathcal{G}_R$ and $\mathcal{G}_C$, which are, respectively, induced by matrices $R \in \mathbb{R}^{m\times m}$ and $C \in \mathbb{R}^{m\times m}$; that is (i, j) is a directed link in the graph $\mathcal{G}_R$ if and only if $R_{ij} > 0$ and, similarly, (i, j) is a directed link in $\mathcal{G}_C$ if and only if $C_{ij} > 0$. We make the following assumption on R and C. (Note that, given a matrix A with non-negative off-diagonal entries, the induced graph does not depend on the diagonal entries of the matrix. Also, $\mathcal{G}_{A^T}$ is identical to $\mathcal{G}_A$ with the directions of edges reversed.)

Assumption 2. *The matrices $R, C \in \mathbb{R}^{m\times m}$ have nonnegative off-diagonal entries ($R_{ij} \geq 0$ and $C_{ij} \geq 0$ for all $i \neq j$). Their diagonal entries are negative, satisfying*

$$R_{ii} = -\sum_{j\in\mathbb{N}_{R,i}^{\text{in}}} R_{ij}, \quad C_{ii} = -\sum_{j\in\mathbb{N}_{C,i}^{\text{out}}} C_{ji} \quad (3)$$

such that R has zero row sums and C has zero column sums. The induced graphs $\mathcal{G}_R$ and $\mathcal{G}_{C^T}$ satisfy:

1) *$\mathcal{G}_R$ is strongly connected, i.e., there is a path (respecting the directions of edges) from each node to every other node;*
2) *The graph induced by C^T, i.e., $\mathcal{G}_{C^T}$, contains at least one spanning tree.*

Remark 1. *The assumption on $\mathcal{G}_{C^T}$ is weaker than requiring that the induced graph $\mathcal{G}_C$ is strongly connected.*

When there is information-sharing noise, shared messages may be corrupted by noise. Namely, when agent i shares x_i^k with agent j, agent j can only receive a distorted version $x_i^k + \zeta_i^k$ of x_i^k, where ζ_i^k denotes the information-sharing noise. Similarly, when agent i shares y_i^k with agent j, agent j can only receive a distorted version $y_i^k + \xi_i^k$ of y_i^k, where ξ_i^k denotes the information-sharing noise. The noises ξ_i^k and ζ_i^k will

significantly impact the accuracy of optimization. In fact, as conventional gradient-tracking algorithms feed the incremental gradient to the y iterate, the noise on y_i^k will accumulate and the variance of noise can grow to infinity as iteration proceeds (this will be detailed later).

To alleviate the influence of information-sharing noise, a decaying factor can be applied to the coupling weight matrix, which has been proven effective in static-consensus based distributed optimization algorithms [20], [35], [36]. However, for gradient-tracking based algorithms, even with a decaying factor on the coupling weight matrices, the noise on y_i^k will still accumulate and increase with time, significantly affecting the accuracy of optimization results. Recently, [38] showed that instead of tracking the global gradient, tracking the cumulative gradient can avoid information noise from accumulating in gradient-tracking based distributed optimization. However, this approach cannot eliminate the influence of persistent information-sharing noise, and it is subject to steady-state errors. Furthermore, it can only avoid noise accumulation when the inter-agent interaction is time-invariant, precluding the possibility of combining a decaying factor (which will make inter-agent interaction time-varying) to gradually attenuate the influence of noise. In this paper, we propose a new algorithm that can achieve both avoidance of noise-accumulation and incorporation of a decaying factor. By sharing the cumulative-gradient estimate (denoted as an s variable) instead of the direct gradient estimate (i.e., the y variable), we can gradually annihilate the influence of information-sharing noise on the estimate of the global gradient, even when the information-sharing noise is persistent.

Algorithm 1: Robust gradient-tracking based distributed optimization

Parameters: Stepsize λ^k and a decaying factor γ^k to suppress information-sharing noise;

Every agent i maintains two states x_i^k and s_i^k, which are initialized randomly with $x_i^0 \in \mathbb{R}^d$ and $s_i^0 \in \mathbb{R}^d$.

for $k = 1, 2, \cdots$ **do**

a) Agent i pushes s_i^k to each agent $l \in \mathbb{N}_{C,i}^{\text{out}}$, which will be received as $s_i^k + \xi_i^k$ due to information-sharing noise. And agent i pulls x_j^k from each $j \in \mathbb{N}_{R,i}^{\text{in}}$, which will be received as $x_j^k + \zeta_j^k$ due to information-sharing noise. Here the subscript R or C in neighbor sets indicates the neighbors with respect to the graphs induced by these matrices.

b) Agent i chooses $\gamma^k > 0$ satisfying $1 + \gamma^k R_{ii} > 0$ and $1 + \gamma^k C_{ii} > 0$ with R_{ii} and C_{ii} defined in (3). Then, agent i updates its states as follows:

$$\begin{aligned} s_i^{k+1} =& (1 + \gamma^k C_{ii}) s_i^k + \gamma^k \sum_{j \in \mathbb{N}_{C,i}^{\text{in}}} C_{ij}(s_j^k + \xi_j^k) \\ &+ \lambda^k \nabla f_i(x_i^k), \\ x_i^{k+1} =& (1 + \gamma^k R_{ii}) x_i^k + \gamma^k \sum_{j \in \mathbb{N}_{R,i}^{\text{in}}} R_{ij}(x_j^k + \zeta_j^k) \\ &- \frac{s_i^{k+1} - s_i^k}{u_i}, \end{aligned} \tag{4}$$

where u_i denotes the ith element of the left eigenvector u^T of $I + \gamma^k R$ associated with eigenvalue 1[1].

c) **end**

Remark 2. *As discussed before, a key difference between the proposed Algorithm 1 and the existing gradient-tracking based algorithms is that Algorithm 1 introduces a decaying factor γ^k to suppress the information-sharing noise. Introducing the decaying factor is reasonable for the following reasons: In the early stages of the iteration, the decaying factor is still far from zero, and hence its attenuation effect on information-sharing is not significant, which allows the necessary mixture of information and hence consensus of individual agents' optimization variables; As the iteration proceeds and individual agents' optimization variables converge to each other (thus diminishing the need for information-sharing), the decaying factor approaches zero and hence its attenuation effect on information-sharing noise becomes more severe, which effectively eliminates the influence of information-sharing noise. Of course, to ensure that necessary gradient descent steps and information-mixture operations can be performed, the decaying factor has to decrease slower than λ^k, which will be specified later in the convergence analysis.*

To compare our algorithm with conventional gradient-tracking based algorithms, we write the algorithm in matrix form. Defining

$$\mathbf{x}^k = \begin{bmatrix} (x_1^k)^T \\ (x_2^k)^T \\ \vdots \\ (x_m^k)^T \end{bmatrix} \in \mathbb{R}^{m \times d}, \ \mathbf{s}^k = \begin{bmatrix} (s_1^k)^T \\ (s_2^k)^T \\ \vdots \\ (s_m^k)^T \end{bmatrix} \in \mathbb{R}^{m \times d},$$

$$\mathbf{g}^k = \begin{bmatrix} (g_1^k)^T \\ (g_2^k)^T \\ \vdots \\ (g_m^k)^T \end{bmatrix} \in \mathbb{R}^{m \times d},$$

with $g_i^k = \nabla f_i(x_i^k)$ and

$$\boldsymbol{\zeta}_w^k = \begin{bmatrix} (\zeta_{w1}^k)^T \\ (\zeta_{w2}^k)^T \\ \vdots \\ (\zeta_{wm}^k)^T \end{bmatrix} \in \mathbb{R}^{m \times d}, \ \boldsymbol{\xi}_w^k = \begin{bmatrix} (\xi_{w1}^k)^T \\ (\xi_{w2}^k)^T \\ \vdots \\ (\xi_{wm}^k)^T \end{bmatrix} \in \mathbb{R}^{m \times d},$$

with

$$\zeta_{wi} \triangleq \sum_{j \in \mathbb{N}_{R,i}^{\text{in}}} R_{ij} \zeta_j^k, \quad \xi_{wi} \triangleq \sum_{j \in \mathbb{N}_{C,i}^{\text{in}}} C_{ij} \xi_j^k,$$

we write the dynamics of (4) in the following more compact form:

$$\begin{aligned} \mathbf{s}^{k+1} &= C^k \mathbf{s}^k + \gamma^k \boldsymbol{\xi}_w^k + \lambda^k \mathbf{g}^k \\ \mathbf{x}^{k+1} &= R^k \mathbf{x}^k + \gamma^k \boldsymbol{\zeta}_w^k - U^{-1}(\mathbf{s}^{k+1} - \mathbf{s}^k) \end{aligned} \tag{5}$$

where

$$R^k = I + \gamma^k R,$$

[1]Under Assumption 2, the matrix $I + \gamma^k R$ always has a unique positive left eigenvector u^T (associated with eigenvalue 1) satisfying $u^T \mathbf{1} = m$ (see details in Lemma 1). When R is balanced, u becomes the vector $\mathbf{1}$ [44].

$$C^k = I + \gamma^k C,$$

and

$$U = \text{diag}(u_1, u_2, \cdots, u_m),$$

with u_i denoting the ith element of R_k's left eigenvector u associated with eigenvalue 1.

It can be seen that in the proposed algorithm, $\mathbf{s}^k - \mathbf{s}^{k-1}$ is fed into the optimization variable and acts as the global-gradient estimate. This new approach will avoid the accumulation of information-sharing noise on the global-gradient estimate, which plagues existing gradient-tracking based approaches. To see this, we use the Push-Pull gradient-tracking algorithm as an example. In the absence of information-sharing noise, the conventional Push-Pull algorithm takes the following form [18]:

$$\begin{aligned} \mathbf{x}^{k+1} &= R^k \mathbf{x}^k - \lambda^k \mathbf{y}^k \\ \mathbf{y}^{k+1} &= C^k \mathbf{y}^k + \mathbf{g}^{k+1} - \mathbf{g}^k. \end{aligned} \tag{6}$$

By setting $\mathbf{y}^0 = \mathbf{g}^0$, one can obtain by induction that

$$\mathbf{1}^T \mathbf{y}^k = \mathbf{1}^T \mathbf{g}^k,$$

i.e., the agents can track the average gradient $\frac{\mathbf{1}^T \mathbf{g}^k}{n}$ by ensuring the consensus of all y_i^k (which leads to $y_i^k = \frac{\mathbf{1}^T \mathbf{y}^k}{m}$ for all i).

However, when exchanged messages are subject to noises, i.e., exchanged x_i^k and y_i^k are received as $x_i^k + \zeta_i^k$ and $y_i^k + \xi_i^k$, respectively, the update of the conventional Push-Pull becomes (after incorporating a decaying factor γ^k)

$$\begin{aligned} \mathbf{x}^{k+1} &= R^k \mathbf{x}^k + \gamma^k \boldsymbol{\zeta}_w^k - \lambda^k \mathbf{y}^k, \\ \mathbf{y}^{k+1} &= C^k \mathbf{y}^k + \gamma^k \boldsymbol{\xi}_w^k + \mathbf{g}^{k+1} - \mathbf{g}^k, \end{aligned} \tag{7}$$

and one can obtain by induction that

$$\mathbf{1}^T \mathbf{y}^k = \mathbf{1}^T \left(\mathbf{g}^k + \sum_{l=0}^{k-1} \gamma^l \boldsymbol{\xi}_w^l \right) \tag{8}$$

even under $\mathbf{y}^0 = \mathbf{g}^0$.

Therefore, under the conventional Push-Pull algorithm, the information-sharing noise accumulates with time (even with a decaying factor γ^k) in the estimate of the global gradient, which significantly compromises optimization accuracy. (This statement is corroborated by the numerical simulation result for the conventional Push-Pull algorithm in [18] in Fig. 1, whose optimization-error variance grows with iterations.) It can be easily verified that other gradient-tracking based distributed optimization algorithms have the same issue of accumulating information-sharing noise.

The proposed algorithm successfully circumvents this problem. In fact, using the update rule of $\mathbf{s}^k$ in (5), one has

$$\begin{aligned} \mathbf{1}^T (\mathbf{s}^{k+1} - \mathbf{s}^k) &= \mathbf{1}^T \left(C^k \mathbf{s}^k + \gamma^k \boldsymbol{\xi}_w^k + \lambda^k \mathbf{g}^k - \mathbf{s}^k \right) \\ &= \mathbf{1}^T \left(\gamma^k C \mathbf{s}^k + \gamma^k \boldsymbol{\xi}_w^k + \lambda^k \mathbf{g}^k \right) \\ &= \mathbf{1}^T \left(\gamma^k \boldsymbol{\xi}_w^k + \lambda^k \mathbf{g}^k \right), \end{aligned} \tag{9}$$

where we used the property $\mathbf{1}^T C = 0$ from the definition of C_{ii} in (3). It is clear that the proposed algorithm avoids information-sharing noise from accumulating on the gradient estimate. It is worth noting that the proposed algorithm achieves avoidance of noise-accumulation even when the inter-agent interaction is time-varying, which enables the incorporation of the decaying factor γ^k and further the final elimination of the influence of information-sharing noise on gradient estimate, even when the noises ζ_i^k and ξ_i^k are persistent. In fact, we can prove that when the decaying factor γ^k is chosen appropriately, the proposed algorithm can guarantee that all agents' x_i^k will converge to the same optimal solution almost surely.

IV. Convergence Analysis

For the convenience of convergence analysis, we first present the following properties for the inter-agent coupling $R^k = I + \gamma^k R$ and $C^k = I + \gamma^k C$:

Lemma 1. *[44] (or Lemma 1 in [18]) Under Assumption 2, for every k, the matrix $I + \gamma^k R$ has a unique positive left eigenvector u^T (associated with eigenvalue 1) satisfying $u^T \mathbf{1} = m$, and the matrix $I + \gamma^k C$ has a unique nonnegative right eigenvector v (associated with eigenvalue 1) satisfying $\mathbf{1}^T v = m$.*

Remark 3. *It is worth noting that the left eigenvector u^T in Lemma 1 is time-invariant and independent of γ^k. In fact, using the definition of left eigenvector, it can be seen that u^T satisfies $u^T(I + \gamma^k R) = u^T$, and thus $u^T(\gamma^k R) = 0$ and further $u^T R = 0$. Namely, u^T corresponds to the left eigenvector of R associated with eigenvalue 0. Given that R has zero row-sums according to Assumption 2, we know that such a u^T always exists. Similarly, we know that the right eigenvector v of $I + \gamma^k C$ is also time-invariant and independent of γ^k.*

According to Lemma 3 in [18], we further know that the spectral radius of $\bar{R}^k \triangleq I + \gamma^k R - \frac{\mathbf{1}u^T}{m}$ is equal to $1 - \gamma^k |\nu_R| < 1$, where ν_R is an eigenvalue of R. Furthermore, there exists a vector norm $\|\mathbf{x}\|_R \triangleq \|\tilde{R}\mathbf{x}\|_2$ (where $\tilde{R}$ is determined by R [18]) such that $\|\bar{R}^k\|_R < 1$ is arbitrarily close to the spectral radius of $\bar{R}^k$, i.e., $1 - \gamma^k |\nu_R| < 1$. Without loss of generality, we represent this norm as $\|\bar{R}^k\|_R = 1 - \gamma^k \rho_R < 1$, where ρ_R is an arbitrarily close approximation of $|\nu_R|$. (Note that for the convergence analysis, we only need the fact that such an $\tilde{R}$ exists, but do not require knowledge of its explicit expression. For an arbitrarily small difference $\epsilon > 0$ between $\|\bar{R}^k\|_R$ and the spectral radius of $\bar{R}^k$, Lemma 5.6.10 in [44] provides a constructive way of finding $\tilde{R}$. Also see Lemma 5 of the extended version of [38] for more discussions about $\tilde{R}$.) Similarly, we have that the spectral radius of $\bar{C}^k \triangleq I + \gamma^k C - \frac{v\mathbf{1}^T}{m}$ is equal to $1 - \gamma^k |\nu_C| < 1$, where ν_C is an eigenvalue of C. Furthermore, there exists a vector norm $\|\mathbf{x}\|_C \triangleq \|\tilde{C}\mathbf{x}\|_2$ (where $\tilde{C}$ is determined by C [18]) such that $\|\bar{C}^k\|_C < 1$ is arbitrarily close to the spectral radius of $\bar{C}^k$, i.e., $1 - \gamma^k |\nu_C| < 1$. Without loss of generality, we represent this norm as $\|\bar{C}^k\|_C = 1 - \gamma^k \rho_c < 1$, where ρ_c is an arbitrarily close approximation of $|\nu_C|$.

For convenience in analysis, we also define the following (weighted) average vectors:

$$\bar{x}^k = \frac{u^T \mathbf{x}^k}{m}, \quad \bar{s}^k = \frac{\mathbf{1}^T \mathbf{s}^k}{m}, \quad \bar{g}^k = \frac{\mathbf{1}^T \mathbf{g}^k}{m}, \tag{10}$$

and

$$\bar{\zeta}_w^k = \frac{u^T \boldsymbol{\zeta}_w^k}{m}, \quad \bar{\xi}_w^k = \frac{\mathbf{1}^T \boldsymbol{\xi}_w^k}{m}. \tag{11}$$

To analyze the convergence of the proposed algorithm, we first present a generic convergence result for gradient-tracking based distributed optimization algorithms. To this end, we first define a matrix norm for $\mathbf{x}^k$ following [18]:

$$\|\mathbf{x}^k\|_R = \left\| \left[\|\mathbf{x}_{(1)}^k\|_R, \|\mathbf{x}_{(2)}^k\|_R, \cdots, \|\mathbf{x}_{(d)}^k\|_R \right] \right\|_2 \tag{12}$$

where the subscript 2 denotes the $2-$norm and $\mathbf{x}_{(i)}^k$ denotes the ith column of $\mathbf{x}^k$ for $1 \leq i \leq d$. One can easily see that $\|\mathbf{x}^k - \mathbf{1}\bar{x}^k\|_R$ measures the distance between all x_i^k and their weighted average $\bar{x}^k$.

Similarly, we define a matrix norm $\|\cdot\|_C$ for $\mathbf{s}^k \triangleq \left[s_1^k, s_2^k, \cdots, s_m^k\right]^T \in \mathbb{R}^{m\times d}$:

$$\|\mathbf{s}^k\|_C = \left\| \left[\|\mathbf{s}_{(1)}^k\|_C, \|\mathbf{s}_{(2)}^k\|_C, \cdots, \|\mathbf{s}_{(d)}^k\|_C \right] \right\|_2 \tag{13}$$

and use $\|\mathbf{s}^k - v\bar{s}^k\|_C$ to measure the distance between all s iterates and their average $\bar{s}^k$ (weighed by v).

We also need the following lemmas about sequences of random vectors:

Lemma 2. *([39], Lemma 4) Let $\{\mathbf{v}^k\} \subset \mathbb{R}^d$ and $\{\mathbf{u}^k\} \subset \mathbb{R}^p$ be random nonnegative vector sequences, and $\{a^k\}$ and $\{b^k\}$ be random nonnegative scalar sequences such that*

$$\mathbb{E}\left[\mathbf{v}^{k+1}|\mathcal{F}^k\right] \leq (V^k + a^k\mathbf{1}\mathbf{1}^T)\mathbf{v}^k + b^k\mathbf{1} - H^k\mathbf{u}^k$$

holds almost surely for all $k \geq 0$, where $\{V^k\}$ and $\{H^k\}$ are random sequences of nonnegative matrices and $\mathbb{E}\left[\mathbf{v}^{k+1}|\mathcal{F}^k\right]$ denotes the conditional expectation given $\mathbf{v}^\ell, \mathbf{u}^\ell, a^\ell, b^\ell, V^\ell, H^\ell$ for $\ell = 0, 1, \ldots, k$. Assume that $\{a^k\}$ and $\{b^k\}$ satisfy $\sum_{k=0}^\infty a^k < \infty$ and $\sum_{k=0}^\infty b^k < \infty$ almost surely, and that there exists a (deterministic) vector $\pi > 0$ such that

$$\pi^T V^k \leq \pi^T, \qquad \pi^T H^k \geq 0, \qquad \forall k \geq 0$$

hold almost surely. Then, we have

1) *$\{\pi^T\mathbf{v}^k\}$ converges almost surely to some random variable $\pi^T\mathbf{v} \geq 0$;*
2) *$\{\mathbf{v}^k\}$ is bounded almost surely;*
3) *$\sum_{k=0}^\infty \pi^T H^k\mathbf{u}^k < \infty$ holds almost surely.*

Lemma 3. *([39], Lemma 7) Let $\{\mathbf{v}^k\} \subset \mathbb{R}^d$ be a sequence of non-negative random vectors and $\{b^k\}$ be a sequence of nonnegative random scalars such that $\sum_{k=0}^\infty b^k < \infty$ and*

$$\mathbb{E}\left[\mathbf{v}^{k+1}|\mathcal{F}^k\right] \leq V^k\mathbf{v}^k + b^k\mathbf{1}, \quad \forall k \geq 0$$

hold almost surely, where $\{V^k\}$ is a sequence of non-negative matrices and $\mathcal{F}^k = \{\mathbf{v}^\ell, b^\ell; 0 \leq \ell \leq k\}$. Assume that there exist a vector $\pi > 0$ and a deterministic scalar sequence $\{a^k\}$ satisfying $a^k \in (0,1)$, $\sum_{k=0}^\infty a^k = \infty$, and $\pi^T V^k \leq (1 - a^k)\pi^T$ for all $k \geq 0$. Then, we have $\lim_{k\to\infty} \mathbf{v}^k = 0$ almost surely.

Now we are in a position to present the generic convergence result for gradient-tracking based distributed optimization algorithms:

Theorem 1. *Assume that the objective function $F(\cdot)$ is continuously differentiable and that the problem (1) has an optimal solution θ^*. Suppose that a distributed algorithm generates a sequence $\{x_i^k\} \subseteq \mathbb{R}^d$ under coupling weight matrix R and a sequence $\{s_i^k\} \subseteq \mathbb{R}^d$ under coupling weight matrix C, such that the following relationship holds almost surely for some sufficiently large integer $T \geq 0$ and for all $k \geq T$:*

$$\mathbb{E}\left[\mathbf{v}^{k+1}|\mathcal{F}^k\right] \leq \left(V^k + a^k\mathbf{1}\mathbf{1}^T\right)\mathbf{v}^k + b^k\mathbf{1} - H^k\begin{bmatrix} \|\nabla F(\bar{x}^k)\|^2 \\ \|\bar{g}^k\|^2 \end{bmatrix} \tag{14}$$

where

$$\mathcal{F}^k = \{x_i^\ell, s_i^\ell; 0 \leq \ell \leq k, \ i \in [m]\},$$

and

$$\mathbf{v}^k = \begin{bmatrix} \mathbf{v}_1^k \\ \mathbf{v}_2^k \\ \mathbf{v}_3^k \end{bmatrix} \triangleq \begin{bmatrix} F(\bar{x}^k) - F(\theta^*) \\ \|\mathbf{x}^k - \mathbf{1}\bar{x}^k\|_R^2 \\ \|\mathbf{s}^k - v\bar{s}^k\|_C^2 \end{bmatrix},$$

$$V^k = \begin{bmatrix} 1 & \kappa_1\lambda^k & 0 \\ 0 & 1-\kappa_2\gamma^k & \kappa_3\gamma^k \\ 0 & 0 & 1-\kappa_4\gamma^k \end{bmatrix},$$

$$H^k = \begin{bmatrix} \kappa_5\lambda^k & \kappa_6\lambda^k - \kappa_7(\lambda^k)^2 \\ 0 & 0 \\ 0 & 0 \end{bmatrix},$$

with $\kappa_i > 0$ for all $1 \leq i \leq 7$ and $\kappa_2, \kappa_4 \in (0,1)$, while the nonnegative scalar sequences $\{a^k\}$, $\{b^k\}$, and positive sequences $\{\lambda^k\}$ and $\{\gamma^k\}$ satisfy $\sum_{k=0}^\infty a^k < \infty$ a.s., $\sum_{k=0}^\infty b^k < \infty$ a.s., $\sum_{k=0}^\infty \lambda^k = \infty$, $\sum_{k=0}^\infty \gamma^k = \infty$, $\sum_{k=0}^\infty (\gamma^k)^2 < \infty$, $\sum_{k=0}^\infty \frac{(\lambda^k)^2}{\gamma^k} < \infty$, and $\lim_{k\to\infty} \lambda^k/\gamma^k = 0$. Then, we have:

(a) $\lim_{k\to\infty} F(\bar{x}^k)$ exists almost surely and

$$\lim_{k\to\infty} \|x_i^k - \bar{x}^k\| = \lim_{k\to\infty} \|s_i^k - v_i\bar{s}^k\| = 0, \ \forall i, \quad a.s.$$

(b) $\liminf_{k\to\infty} \|\nabla F(\bar{x}^k)\| = 0$ holds almost surely. Moreover, if the function $F(\cdot)$ has bounded level sets, then $\{\bar{x}^k\}$ is bounded and every accumulation point of $\{\bar{x}^k\}$ is an optimal solution almost surely, and

$$\lim_{k\to\infty} F(x_i^k) = F(\theta^*), \qquad \forall i \in [m], \quad a.s.$$

Proof. Since the results of Lemma 2 are asymptotic, they remain valid when the starting index is shifted from $k = 0$ to $k = T$, for an arbitrary $T \geq 0$. So the idea is to show that the conditions in Lemma 2 are satisfied for all $k \geq T$, where $T \geq 0$ is large enough.

(a) Because $\kappa_i > 0$ for all $1 \leq i \leq 7$, for $\pi = [\pi_1, \pi_2, \pi_3]^T$ to satisfy $\pi^T V \leq \pi^T$ and $\pi^T H^k \geq 0$, we only need to show that the following inequalities are true

$$\begin{aligned} &\kappa_1\lambda^k\pi_1 + (1-\kappa_2\gamma^k)\pi_2 \leq \pi_2, \\ &\kappa_3\gamma^k\pi_2 + (1-\kappa_4\gamma^k)\pi_3 \leq \pi_3, \\ &\left(\kappa_6\lambda^k - \kappa_7(\lambda^k)^2\right)\pi_1 \geq 0. \end{aligned} \tag{15}$$

The first inequality is equivalent to $\pi_2 \geq \frac{\kappa_1\lambda^k}{\kappa_2\gamma^k}\pi_1$. Given that $\lim_{k\to\infty} \lambda^k/\gamma^k = 0$ holds and γ^k as well as λ^k is positive according to the assumption, it can easily be seen that for any given $\pi_1 > 0$, we can always find a $\pi_2 > 0$ satisfying the relationship when k is larger than some $T \geq 0$.

The second inequality is equivalent to $\pi_3 \geq \frac{\kappa_3}{\kappa_4}\pi_2$, which can always be satisfied by setting $\pi_3 = \frac{\kappa_3}{\kappa_4}\pi_2$ after fixing π_2.

The third inequality is equivalent to $\kappa_6 - \kappa_7\lambda^k > 0$, which is always satisfied given that $(\lambda^k)^2$ is summable (and hence λ^k tends to zero).

Thus, we can always find a vector π satisfying all inequalities in (15) for $k \geq T$ for some large enough $T \geq 0$, and hence the conditions in Lemma 2 are satisfied.

By Lemma 2, it follows that for the three entries of $\mathbf{v}^k$, i.e., $\mathbf{v}_1^k$, $\mathbf{v}_2^k$, and $\mathbf{v}_3^k$, we have that

$$\lim_{k\to\infty} \pi_1\mathbf{v}_1^k + \pi_2\mathbf{v}_2^k + \pi_3\mathbf{v}_3^k \tag{16}$$

exists almost surely, and

$$\sum_{k=0}^{\infty} \pi^T H^k \mathbf{u}^k < \infty$$

holds almost surely with

$$\mathbf{u}^k = [\|\nabla F(\bar{x}^k)\|^2,\ \|\bar{g}^k\|^2]^T.$$

Since $\pi^T H^k$ has the following form

$$\pi^T H^k = \left[\kappa_5\lambda^k\pi_1,\ (\kappa_6\lambda^k - \kappa_7(\lambda^k)^2)\pi_1\right]$$

and $(\lambda^k)^2$ is summable, one has

$$\sum_{k=0}^{\infty} \lambda^k\|\nabla F(\bar{x}^k)\|^2 < \infty, \quad \sum_{k=0}^{\infty} \lambda^k\|\bar{g}^k\|^2 < \infty, \quad a.s. \tag{17}$$

Hence, it follows that

$$\|\nabla F(\bar{x}^k)\| < \Delta_1, \quad \|\bar{g}^k\| < \Delta_2, \quad a.s. \tag{18}$$

for some random scalars $\Delta_1 > 0$ and $\Delta_2 > 0$ due to the assumption $\sum_{k=0}^{\infty}\lambda^k = \infty$.

Now, we focus on proving that both $\mathbf{v}_2^k = \|\mathbf{x}^k - \mathbf{1}\bar{x}^k\|_R^2$ and $\mathbf{v}_3^k = \|\mathbf{s}^k - v\bar{s}^k\|_C^2$ converge to 0 almost surely. The idea is to show that we can apply Lemma 3. By focusing on the second and third elements of $\mathbf{v}^k$, i.e., $\mathbf{v}_2^k$ and $\mathbf{v}_3^k$, from (14) we have

$$\begin{bmatrix} \mathbf{v}_2^{k+1} \\ \mathbf{v}_3^{k+1} \end{bmatrix} \leq \left(\tilde{V}^k + a^k\mathbf{1}\mathbf{1}^T\right)\begin{bmatrix} \mathbf{v}_2^k \\ \mathbf{v}_3^k \end{bmatrix} + \hat{b}^k\mathbf{1},$$

where

$$\hat{b}^k = b^k + a^k(F(\bar{x}^k) - F(\theta^*)),$$

$$\tilde{V}^k = \begin{bmatrix} 1-\kappa_2\gamma^k & \kappa_3\gamma^k \\ 0 & 1-\kappa_4\gamma^k \end{bmatrix},$$

which can be rewritten as

$$\begin{bmatrix} \mathbf{v}_2^{k+1} \\ \mathbf{v}_3^{k+1} \end{bmatrix} \leq \tilde{V}^k\begin{bmatrix} \mathbf{v}_2^k \\ \mathbf{v}_3^k \end{bmatrix} + \tilde{b}^k\mathbf{1} \tag{19}$$

with

$$\tilde{b}^k = b^k + a^k\left(F(\bar{x}^k) - F(\theta^*) + \|\mathbf{x}^k - \mathbf{1}\bar{x}^k\|_R^2 + \|\mathbf{s}^k - v\bar{s}^k\|_C^2\right).$$

To apply Lemma 3, noting that γ^k is not summable, we show that the inequality $\tilde{\pi}^T\tilde{V}^k \leq (1-\alpha\gamma^k)\tilde{\pi}^T$ has a solution in $\tilde{\pi} = [\pi_2, \pi_3]$ with $\pi_2, \pi_3 > 0$ and $\alpha \in (0,1)$.

From

$$\tilde{\pi}^T\tilde{V}^k \leq (1-\alpha\gamma^k)\tilde{\pi}^T,$$

one has

$$(1-\kappa_2\gamma^k)\pi_2 \leq (1-\alpha\gamma^k)\pi_2$$

and

$$\kappa_3\gamma^k\pi_2 + (1-\kappa_4\gamma^k)\pi_3 \leq (1-\alpha\gamma^k)\pi_3,$$

which can be simplified as $\alpha \leq \kappa_2$ and $\alpha \leq \kappa_4 - \frac{\pi_2}{\pi_3}\kappa_3$.

Given $\kappa_2 > 0$, $\kappa_3 > 0$, and $\kappa_4 > 0$ according to our assumption, we can always find appropriate $\pi_2 > 0$ and $\pi_3 > 0$ to make $\alpha \in (0,1)$ hold.

We next prove that the condition $\sum_{k=0}^{\infty}\tilde{b}^k < 0$ *a.s.* of Lemma 3 is also satisfied. Indeed, the condition can be met because: (1) b^k and a^k are summable according to the assumption of the theorem; and (2) $F(\bar{x}^k) - F(\theta^*)$, $\|\mathbf{x}^k - \mathbf{1}\bar{x}^k\|_R^2$, $\|\mathbf{s}^k - v\bar{s}^k\|_C^2$ are all bounded almost surely due to the existence of the limit in (16). Thus, all the conditions of Lemma 3 are satisfied, and thus it follows that $\lim_{k\to\infty}\|x_i^k - \bar{x}^k\| = 0$ and $\lim_{k\to\infty}\|s_i^k - v_i\bar{s}^k\| = 0$ hold almost surely. Moreover, in view of the existence of the limit in (16) and the facts that $\pi_1 > 0$ and $v_1^k = F(\bar{x}^k) - F(\theta^*)$, it follows that $\lim_{k\to\infty}F(\bar{x}^k)$ exists almost surely.

(b) Since $\sum_{k=0}^{\infty}\lambda^k\|\nabla F(\bar{x}^k)\|^2 < \infty$ holds almost surely (see (17)), from $\sum_{k=0}^{\infty}\lambda^k = \infty$, it follows that we have $\liminf_{k\to\infty}\|\nabla F(\bar{x}^k)\| = 0$ almost surely.

Now, if the function $F(\cdot)$ has bounded level sets, then the sequence $\{\bar{x}^k\}$ is bounded almost surely since the limit $\lim_{k\to\infty}F(\bar{x}^k)$ exists almost surely (as shown in part (a)). Thus, $\{\bar{x}^k\}$ has accumulation points almost surely. Let $\{\bar{x}^{k_i}\}$ be a sub-sequence such that $\lim_{i\to\infty}\|\nabla F(\bar{x}^{k_i})\| = 0$ holds almost surely. Without loss of generality, we may assume that $\{\bar{x}^{k_i}\}$ is almost surely convergent, for otherwise we would choose a sub-sequence of $\{\bar{x}^{k_i}\}$. Let $\lim_{i\to\infty}\bar{x}^{k_i} = \hat{x}$. Then, by the continuity of the gradient $\nabla F(\cdot)$, it follows that $\nabla F(\hat{x}) = 0$, implying that $\hat{x}$ is an optimal point. Since $F(\cdot)$ is continuous, we have $\lim_{i\to\infty}F(\bar{x}^{k_i}) = F(\hat{x}) = F(\theta^*)$. By part (a), $\lim_{k\to\infty}F(\bar{x}^k)$ exists almost surely, and thus we must have $\lim_{k\to\infty}F(\bar{x}^k) = F(\theta^*)$ almost surely.

Finally, by part (a), we have $\lim_{k\to\infty}\|x_i^k - \bar{x}^k\|^2 = 0$ almost surely for every i. Thus, each $\{x_i^k\}$ has the same accumulation points as the sequence $\{\bar{x}^k\}$ almost surely, implying by the continuity of the function $F(\cdot)$ that $\lim_{k\to\infty}F(x_i^k) = F(\theta^*)$ holds almost surely for all i. ■

Remark 4. *In Theorem 1(b), the bounded level set condition can be replaced with any other condition ensuring that the sequence $\{\bar{x}^k\}$ is bounded almost surely.*

Theorem 1 is critical for establishing convergence properties of the gradient tracking-based distributed algorithm together with suitable conditions on the information-sharing noise. We make the following assumption on the noise:

Assumption 3. *For every $i \in [m]$, the noise sequences $\{\zeta_i^k\}$ and $\{\xi_i^k\}$ are zero-mean independent random variables, and independent of $\{x_i^0; i \in [m]\}$. Also, for every k, the noise collection $\{\zeta_j^k, \xi_j^k; j \in [m]\}$ is independent. The noise*

variances $(\sigma_{\zeta,i}^k)^2 = \mathbb{E}\left[\|\zeta_i^k\|^2\right]$ *and* $(\sigma_{\xi,i}^k)^2 = \mathbb{E}\left[\|\xi_i^k\|^2\right]$ *and the decaying factor* γ^k *are such that*

$$\sum_{k=0}^{\infty}(\gamma^k)^2 \max_{i\in[m]}(\sigma_{\zeta,i}^k)^2 < \infty, \sum_{k=0}^{\infty}(\gamma^k)^2 \max_{j\in[m]}(\sigma_{\xi,j}^k)^2 < \infty. \tag{20}$$

The initial random vectors satisfy $\mathbb{E}\left[\|x_i^0\|^2\right] < \infty$, $\forall i \in [m]$.

Remark 5. *The condition (20) is satisfied, for example, when sequences* $\{(\gamma^k)^2\}$ *and* $\{(\lambda^k)^2\}$ *are summable, and sequences* $\{\sigma_{\zeta,i}^k\}$ *and* $\{\sigma_{\xi,i}^k\}$ *are bounded for every* $i \in [m]$.

Theorem 2. *Let Assumption 1, Assumption 2, and Assumption 3 hold. If* $\{\gamma^k\}$ *and* $\{\lambda^k\}$ *satisfy* $\sum_{k=0}^{\infty}\gamma^k = \infty$, $\sum_{k=0}^{\infty}(\gamma^k)^2 < \infty$, $\sum_{k=0}^{\infty}\lambda^k = \infty$, $\sum_{k=0}^{\infty}\frac{(\lambda^k)^2}{\gamma^k} < \infty$, *and* $\lim_{k\to\infty}\lambda^k/\gamma^k = 0$, *then the results of Theorem 1 hold for Algorithm 1.*

Proof. The goal is to establish the relationship in (14), with the σ-field $\mathcal{F}^k = \{x_i^\ell, s_i^\ell; 0 \le \ell \le k,\ i \in [m]\}$. To this end, we divide the derivations into four steps: in Step I, Step II, and Step III, we establish relations for $\|\mathbf{s}^k - v\bar{s}^k\|_C$, $\|\mathbf{x}^k - \mathbf{1}\bar{x}^k\|_R$, and $\mathbb{E}\left[F(\bar{x}^k) - F(\theta^*)|\mathcal{F}^k\right]$ for the iterates generated by the proposed algorithm, respectively. In Step IV, we use them to show that (14) of Theorem 1 holds.

Step I: Relationship for $\|\mathbf{s}^k - v\bar{s}^k\|_C$.

From (5), we have

$$\begin{aligned}\bar{s}^{k+1} &= \frac{\mathbf{1}^T\mathbf{s}^{k+1}}{m}\\ &= \frac{\mathbf{1}^T}{m}\left(C^k\mathbf{s}^k + \gamma^k\boldsymbol{\xi}_w^k + \lambda^k\mathbf{g}^k\right)\\ &= \bar{s}^k + \gamma^k\bar{\xi}_w^k + \lambda^k\bar{g}^k,\end{aligned} \tag{21}$$

which, in combination with the relationship $\left(C^k - \frac{v\mathbf{1}^T}{m}\right)v = 0$, leads to

$$\mathbf{s}^{k+1} - v\bar{s}^{k+1} = \bar{C}^k(\mathbf{s}^k - v\bar{s}^k) + \gamma^k\Pi_v\boldsymbol{\xi}_w^k + \lambda^k\Pi_v\mathbf{g}^k,$$

where we used the relationship $\bar{C}^k = C^k - \frac{v\mathbf{1}^T}{m}$ and defined $\Pi_v = I - \frac{v\mathbf{1}^T}{m}$ for the sake of notational simplicity.

The preceding relationship further leads to

$$\begin{aligned}&\left\|\mathbf{s}^{k+1} - v\bar{s}^{k+1}\right\|_C^2\\ &= \left\|\bar{C}^k(\mathbf{s}^k - v\bar{s}^k) + \lambda^k\Pi_v\mathbf{g}^k\right\|_C^2 + \left\|\gamma^k\Pi_v\boldsymbol{\xi}_w^k\right\|_C^2\\ &\quad + 2\left\langle\bar{C}^k(\mathbf{s}^k - v\bar{s}^k) + \lambda^k\Pi_v\mathbf{g}^k, \gamma^k\Pi_v\boldsymbol{\xi}_w^k\right\rangle_C\\ &\le \left(\|\bar{C}^k\|_C\|\mathbf{s}^k - v\bar{s}^k\|_C + \lambda^k\|\Pi_v\|_C\|\mathbf{g}^k\|_C\right)^2\\ &\quad + \left\|\gamma^k\Pi_v\boldsymbol{\xi}_w^k\right\|_C^2\\ &\quad + 2\left\langle\bar{C}^k(\mathbf{s}^k - v\bar{s}^k) + \lambda^k\Pi_v\mathbf{g}^k, \gamma^k\Pi_v\boldsymbol{\xi}_w^k\right\rangle_C,\end{aligned} \tag{22}$$

where $\langle\cdot\rangle_C$ denotes the inner product induced[2] by the norm $\|\cdot\|_C$.

We further bound the first term on the right hand side of the preceding inequality using the property $\|\bar{C}^k\|_C = 1 - \gamma^k\rho_c$ and the inequality $(a+b)^2 \le (1+\epsilon)a^2 + (1+\epsilon^{-1})b^2$ valid for any scalars a, b, and $\epsilon > 0$ (by setting $\epsilon = \frac{1}{1-\gamma^k\rho_c} - 1$ and hence $1 - \epsilon^{-1} = \frac{1}{\gamma^k\rho_c}$):

$$\begin{aligned}&\left\|\mathbf{s}^{k+1} - v\bar{s}^{k+1}\right\|_C^2\\ &\le \left((1-\gamma^k\rho_c)\|\mathbf{s}^k - v\bar{s}^k\|_C + \lambda^k\|\Pi_v\|_C\|\mathbf{g}^k\|_C\right)^2\\ &\quad + \left\|\gamma^k\Pi_v\boldsymbol{\xi}_w^k\right\|_C^2\\ &\quad + 2\left\langle\bar{C}^k(\mathbf{s}^k - v\bar{s}^k) + \lambda^k\Pi_v\mathbf{g}^k, \gamma^k\Pi_v\boldsymbol{\xi}_w^k\right\rangle_C\\ &\le (1-\gamma^k\rho_c)\|\mathbf{s}^k - v\bar{s}^k\|_C^2\\ &\quad + \frac{(\lambda^k)^2}{\gamma^k\rho_c}\|\Pi_v\|_C^2\|\mathbf{g}^k\|_C^2 + \left\|\gamma^k\Pi_v\boldsymbol{\xi}_w^k\right\|_C^2\\ &\quad + 2\left\langle\bar{C}^k(\mathbf{s}^k - v\bar{s}^k) + \lambda^k\Pi_v\mathbf{g}^k, \gamma^k\Pi_v\boldsymbol{\xi}_w^k\right\rangle_C.\end{aligned}$$

Taking the expectation (conditioned on $\mathcal{F}^k$) on both sides yields

$$\begin{aligned}&\mathbb{E}\left[\left\|\mathbf{s}^{k+1} - v\bar{s}^{k+1}\right\|_C^2|\mathcal{F}^k\right] \le (1-\gamma^k\rho_c)\left\|\mathbf{s}^k - v\bar{s}^k\right\|_C^2\\ &\quad + \frac{(\lambda^k)^2}{\gamma^k\rho_c}\|\Pi_v\|_C^2\|\mathbf{g}^k\|_C^2 + (\gamma^k)^2\|\Pi_v\|_C^2\mathbb{E}\left[\|\boldsymbol{\xi}_w^k\|_C^2\right]\\ &\le (1-\gamma^k\rho_c)\left\|\mathbf{s}^k - v\bar{s}^k\right\|_C^2 + \frac{(\lambda^k)^2\delta_{C,2}^2}{\gamma^k\rho_c}\|\Pi_v\|_C^2\|\mathbf{g}^k\|_2^2\\ &\quad + (\gamma^k)^2\delta_{C,2}^2\|\Pi_v\|_C^2\mathbb{E}\left[\|\boldsymbol{\xi}_w^k\|_2^2\right]\\ &= (1-\gamma^k\rho_c)\left\|\mathbf{s}^k - v\bar{s}^k\right\|_C^2 + \frac{(\lambda^k)^2\delta_{C,2}^2}{\gamma^k\rho_c}\|\Pi_v\|_C^2\|\mathbf{g}^k\|_2^2\\ &\quad + (\gamma^k)^2\delta_{C,2}^2\|\Pi_v\|_C^2\sum_{i,j}(C_{ij}\sigma_{\xi,j}^k)^2,\end{aligned} \tag{23}$$

where $\delta_{C,2}$ is a constant such that $\|x\|_C \le \delta_{C,2}\|x\|_2$ for all x. (In finite-dimensional vector spaces, all norms are equivalent up to a proportionality constant, represented by $\delta_{C,2}$ here.) Note that the inner-product term in the preceding step disappears because the means of all ξ_i^k are zero according to Assumption 3, and hence their linear combination $\boldsymbol{\xi}_w^k$ also has zero mean.

Next we proceed to bound the term $\|\mathbf{g}^k\|_2$ on the right hand side of the preceding inequality.

Because every $f_i(\cdot)$ is convex with Lipschitz continuous gradient L according to Assumption 1, we always have the following relation (see Theorem 2.1.5 in [45]):

$$f_i(v) + \langle\nabla f_i(v), u - v\rangle + \frac{\|\nabla f_i(v) - \nabla f_i(u)\|^2}{2L} \le f_i(u)$$

for any $u, v \in \mathbb{R}^d$.

Letting $v = \theta^*$ and $u = \bar{x}^k$ in the preceding relation, we obtain for all i

$$f_i(\theta^*) + \left\langle\nabla f_i(\theta^*), \bar{x}^k - \theta^*\right\rangle + \frac{\|\nabla f_i(\theta^*) - \nabla f_i(\bar{x}^k)\|^2}{2L} \le f_i(\bar{x}^k)$$

and further

$$F(\theta^*) + \left\langle\nabla F(\theta^*), \bar{x}^k - \theta^*\right\rangle + \frac{\sum_{i=1}^m\|\nabla f_i(\theta^*) - \nabla f_i(\bar{x}^k)\|^2}{2mL} \le F(\bar{x}^k).$$

[2]Since one can verify that $\|\mathbf{s}^k\|_C = \|\tilde{C}\mathbf{s}^k\|_2$ where $\tilde{C}$ is discussed in the paragraph after Remark 3, we have the norm $\|\cdot\|_C$ satisfying the Parallelogram law, implying that it has an associated inner product $\langle\cdot,\cdot\rangle_C$.

Recalling $\nabla F(\theta^*)=0$, we have

$$\sum_{i=1}^m \|\nabla f_i(\theta^*)-\nabla f_i(\bar{x}^k)\|^2 \le 2mL(F(\bar{x}^k)-F(\theta^*))$$

and further

$$\begin{aligned}
&\sum_{i=1}^m \|\nabla f_i(\bar{x}^k)\|^2\\
&\le 2\sum_{i=1}^m \left(\|\nabla f_i(\theta^*)-\nabla f_i(\bar{x}^k)\|^2+\|\nabla f_i(\theta^*)\|^2\right) \qquad (24)\\
&\le 4mL(F(\bar{x}^k)-F(\theta^*))+2\sum_{i=1}^m \|\nabla f_i(\theta^*)\|^2.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\|\mathbf{g}^k\|^2 &= \sum_{i=1}^m \|g_i^k\|^2\\
&\le 2\sum_{i=1}^m \left(\|g_i^k-\nabla f_i(\bar{x}^k)\|^2+\|\nabla f_i(\bar{x}^k)\|^2\right)\\
&\le 2L^2\sum_{i=1}^m \|x_i^k-\bar{x}^k\|^2+8mL(F(\bar{x}^k)-F(\theta^*)) \qquad (25)\\
&\quad +4\sum_{i=1}^m \|\nabla f_i(\theta^*)\|^2\\
&= 2L^2\|\mathbf{x}^k-\mathbf{1}\bar{x}^k\|_2^2+8mL(F(\bar{x}^k)-F(\theta^*))\\
&\quad +4\sum_{i=1}^m \|\nabla f_i(\theta^*)\|^2.
\end{aligned}$$

Plugging (25) into (23) yields

$$\begin{aligned}
\mathbb{E}\left[\left\|\mathbf{s}^{k+1}-v\bar{s}^{k+1}\right\|_C^2|\mathcal{F}^k\right] &\le (1-\gamma^k\rho_c)\left\|\mathbf{s}^k-v\bar{s}^k\right\|_C^2\\
&+\frac{2L^2(\lambda^k)^2\delta_{C,2}^2\|\Pi_v\|_C^2}{\gamma^k\rho_c}\|\mathbf{x}^k-\mathbf{1}\bar{x}^k\|_2^2\\
&+\frac{8mL(\lambda^k)^2\delta_{C,2}^2\|\Pi_v\|_C^2}{\gamma^k\rho_c}(F(\bar{x}^k)-F(\theta^*)) \qquad (26)\\
&+\frac{4(\lambda^k)^2\delta_{C,2}^2\|\Pi_v\|_C^2}{\gamma^k\rho_c}\sum_{i=1}^m\|\nabla f_i(\theta^*)\|^2\\
&+(\gamma^k)^2\delta_{C,2}^2\|\Pi_v\|_C^2\sum_{i,j}(C_{ij}\sigma_{\xi,j}^k)^2.
\end{aligned}$$

Step II: Relationship for $\|\mathbf{x}^k-\mathbf{1}\bar{x}^k\|_R$.

From (5), we obtain

$$\begin{aligned}
\bar{x}^{k+1} &= \frac{u^T}{m}\mathbf{x}^{k+1}=\frac{u^T}{m}(R^k\mathbf{x}^k+\gamma^k\boldsymbol{\zeta}_w^k-U^{-1}(\mathbf{s}^{k+1}-\mathbf{s}^k))\\
&= \bar{x}^k+\gamma^k\bar{\zeta}_w^k-\frac{\mathbf{1}^T}{m}(\mathbf{s}^{k+1}-\mathbf{s}^k)\\
&= \bar{x}^k+\gamma^k\bar{\zeta}_w^k-(\bar{s}^{k+1}-\bar{s}^k)\\
&= \bar{x}^k+\gamma^k\bar{\zeta}_w^k-\gamma^k\bar{\xi}_w^k-\lambda^k\bar{g}^k,
\end{aligned} \qquad (27)$$

where we used $u^TU^{-1}=\mathbf{1}^T$ in the second equality and (21) in the last equality.

Combining (5) and (27) leads to

$$\mathbf{x}^{k+1}-\mathbf{1}\bar{x}^{k+1}=\bar{R}^k(\mathbf{x}^k-\mathbf{1}\bar{x}^k)-\Pi_U(\mathbf{s}^{k+1}-\mathbf{s}^k)+\gamma^k\Pi_u\boldsymbol{\zeta}_w^k, \qquad (28)$$

where we used the relationship $\bar{R}^k\mathbf{1}\bar{x}^k=0$ and $\bar{R}^k=R^k-\frac{\mathbf{1}u^T}{m}$, and defined $\Pi_u=I-\frac{\mathbf{1}u^T}{m}$, $\Pi_U=U^{-1}-\frac{\mathbf{1}\mathbf{1}^T}{m}$ for the sake of notational simplicity.

From the first relationship in (5), we can obtain

$$\begin{aligned}
\mathbf{s}^{k+1}-\mathbf{s}^k &= C^k\mathbf{s}^k+\gamma^k\boldsymbol{\xi}_w^k+\lambda^k\mathbf{g}^k-\mathbf{s}^k\\
&= \gamma^kC\mathbf{s}^k+\gamma^k\boldsymbol{\xi}_w^k+\lambda^k\mathbf{g}^k \qquad (29)\\
&= \gamma^kC(\mathbf{s}^k-v\bar{s}^k)+\gamma^k\boldsymbol{\xi}_w^k+\lambda^k\mathbf{g}^k,
\end{aligned}$$

where we used $C^k=I+\gamma^kC$ in the second equality and $Cv=0$ in the last equality.

Combining (28) and (29) yields

$$\begin{aligned}
\mathbf{x}^{k+1}-\mathbf{1}\bar{x}^{k+1} =&\bar{R}^k(\mathbf{x}^k-\mathbf{1}\bar{x}^k)-\gamma^k\Pi_UC(\mathbf{s}^k-v\bar{s}^k)\\
&-\gamma^k\Pi_U\boldsymbol{\xi}_w^k-\lambda^k\Pi_U\mathbf{g}^k+\gamma^k\Pi_u\boldsymbol{\zeta}_w^k.
\end{aligned} \qquad (30)$$

Taking the norm $\|\cdot\|_R$ on both sides of the preceding relationship yields

$$\begin{aligned}
&\|\mathbf{x}^{k+1}-\mathbf{1}\bar{x}^{k+1}\|_R^2\\
&=\|\bar{R}^k(\mathbf{x}^k-\mathbf{1}\bar{x}^k)-\gamma^k\Pi_UC(\mathbf{s}^k-v\bar{s}^k)-\lambda^k\Pi_U\mathbf{g}^k\|_R^2\\
&\quad+(\gamma^k)^2\|\Pi_u\boldsymbol{\zeta}_w^k-\Pi_U\boldsymbol{\xi}_w^k\|_R^2\\
&\quad+2\left\langle\bar{R}^k(\mathbf{x}^k-\mathbf{1}\bar{x}^k)-\gamma^k\Pi_UC(\mathbf{s}^k-v\bar{s}^k)-\lambda^k\Pi_U\mathbf{g}^k,\right.\\
&\qquad\left.\gamma^k\Pi_u\boldsymbol{\zeta}_w^k-\gamma^k\Pi_U\boldsymbol{\xi}_w^k\right\rangle_R\\
&\le\left(\|\bar{R}^k\|_R\|\mathbf{x}^k-\mathbf{1}\bar{x}^k\|_R+\gamma^k\|\Pi_UC\|_R\|\mathbf{s}^k-v\bar{s}^k\|_R\right.\\
&\quad\left.+\lambda^k\|\Pi_U\|_R\|\mathbf{g}^k\|_R\right)^2+(\gamma^k)^2\|\Pi_u\boldsymbol{\zeta}_w^k-\Pi_U\boldsymbol{\xi}_w^k\|_R^2\\
&\quad+2\left\langle\bar{R}^k(\mathbf{x}^k-\mathbf{1}\bar{x}^k)-\gamma^k\Pi_UC(\mathbf{s}^k-v\bar{s}^k)-\lambda^k\Pi_U\mathbf{g}^k,\right.\\
&\qquad\left.\gamma^k\Pi_u\boldsymbol{\zeta}_w^k-\gamma^k\Pi_U\boldsymbol{\xi}_w^k\right\rangle_R,
\end{aligned} \qquad (31)$$

where $\langle\cdot\rangle_R$ denotes the inner product induced[3] by the norm $\|\cdot\|_R$.

Using the relationship $\|\bar{R}^k\|_R=1-\gamma^k\rho_R$ and the inequality $(a+b)^2\le(1+\epsilon)a^2+(1+\epsilon^{-1})b^2$ valid for any scalars a, b, and $\epsilon>0$ (by setting $\epsilon=\frac{1}{1-\gamma^k\rho_R}-1$ and hence $1-\epsilon^{-1}=\frac{1}{\gamma^k\rho_R}$), we can arrive at

$$\begin{aligned}
&\|\mathbf{x}^{k+1}-\mathbf{1}\bar{x}^{k+1}\|_R^2\le(1-\gamma^k\rho_R)\|\mathbf{x}^k-\mathbf{1}\bar{x}^k\|_R^2\\
&\quad+\frac{2\gamma^k\|\Pi_UC\|_R^2}{\rho_R}\|\mathbf{s}^k-v\bar{s}^k\|_R^2+\frac{2(\lambda^k)^2\|\Pi_U\|_R^2}{\gamma^k\rho_R}\|\mathbf{g}^k\|_R^2\\
&\quad+(\gamma^k)^2\|\Pi_u\boldsymbol{\zeta}_w^k-\Pi_U\boldsymbol{\xi}_w^k\|_R^2\\
&\quad+2\left\langle\bar{R}^k(\mathbf{x}^k-\mathbf{1}\bar{x}^k)-\gamma^k\Pi_UC(\mathbf{s}^k-v\bar{s}^k)-\lambda^k\Pi_U\mathbf{g}^k,\right.\\
&\qquad\left.\gamma^k\Pi_u\boldsymbol{\zeta}_w^k-\gamma^k\Pi_U\boldsymbol{\xi}_w^k\right\rangle_R.
\end{aligned} \qquad (32)$$

[3]Since one can verify that $\|\mathbf{x}^k\|_R=\|\tilde{R}\mathbf{x}^k\|_2$ where $\tilde{R}$ is discussed in the paragraph after Remark 3, we have the norm $\|\cdot\|_R$ satisfying the Parallelogram law, implying that it has an associated inner product $\langle\cdot,\cdot\rangle_R$.

Taking the expectation (conditioned on $\mathcal{F}^k$) on both sides yields

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{x}^{k+1}-\mathbf{1}\bar{x}^{k+1}\|_R^2|\mathcal{F}^k\right] \leq (1-\gamma^k\rho_R)\|\mathbf{x}^k-\mathbf{1}\bar{x}^k\|_R^2 \\
&\quad+\frac{2\gamma^k\|\Pi_U C\|_R^2}{\rho_R}\|\mathbf{s}^k-v\bar{s}^k\|_R^2+\frac{2(\lambda^k)^2\|\Pi_U\|_R^2}{\gamma^k\rho_R}\|\mathbf{g}^k\|_R^2 \\
&\quad+2(\gamma^k)^2\|\Pi_u\|^2\mathbb{E}\left[\|\boldsymbol{\zeta}_w^k\|_R^2\right]+2(\gamma^k)^2\|\Pi_U\|_R^2\mathbb{E}\left[\|\boldsymbol{\xi}_w^k\|_R^2\right] \\
&\leq (1-\gamma^k\rho_R)\|\mathbf{x}^k-\mathbf{1}\bar{x}^k\|_R^2+\frac{2\gamma^k\|\Pi_U C\|_R^2}{\rho_R}\|\mathbf{s}^k-v\bar{s}^k\|_R^2 \\
&\quad+\frac{2(\lambda^k)^2\|\Pi_U\|_R^2\delta_{R,2}^2}{\gamma^k\rho_R}\|\mathbf{g}^k\|_2^2 \\
&\quad+2(\gamma^k)^2\|\Pi_u\|_R^2\delta_{R,2}^2\sum_{i,j}(R_{ij}\sigma_{\zeta,j}^k)^2 \\
&\quad+2(\gamma^k)^2\|\Pi_U\|_R^2\delta_{R,2}^2\sum_{i,j}(C_{ij}\sigma_{\xi,j}^k)^2,
\end{aligned} \tag{33}$$

where $\delta_{R,2}$ is a constant such that $\|x\|_R \leq \delta_{R,2}\|x\|_2$ for all x. (As mentioned earlier, in finite-dimensional vector spaces, all norms are equivalent up to a proportionality constant, represented here by $\delta_{R,2}$.)

Plugging (25) into (33) yields

$$\begin{aligned}
&\mathbb{E}\left[\|\mathbf{x}^{k+1}-\mathbf{1}\bar{x}^{k+1}\|_R^2|\mathcal{F}^k\right] \\
&\leq\left(1-\gamma^k\rho_R+\frac{4(\lambda^k)^2L^2\|\Pi_U\|_R^2\delta_{R,2}^2}{\gamma^k\rho_R}\right)\|\mathbf{x}^k-\mathbf{1}\bar{x}^k\|_R^2 \\
&\quad+\frac{2\gamma^k\|\Pi_U C\|_R^2}{\rho_R}\|\mathbf{s}^k-v\bar{s}^k\|_R^2 \\
&\quad+\frac{16mL(\lambda^k)^2\|\Pi_U\|_R^2\delta_{R,2}^2}{\gamma^k\rho_R}(F(\bar{x}^k)-F(\theta^*)) \\
&\quad+\frac{8(\lambda^k)^2\|\Pi_U\|_R^2\delta_{R,2}^2}{\gamma^k\rho_R}\sum_{i=1}^m\|\nabla f_i(\theta^*)\|^2 \\
&\quad+2(\gamma^k)^2\|\Pi_u\|_R^2\delta_{R,2}^2\sum_{i,j}(R_{ij}\sigma_{\zeta,j}^k)^2 \\
&\quad+2(\gamma^k)^2\|\Pi_U\|_R^2\delta_{R,2}^2\sum_{i,j}(C_{ij}\sigma_{\xi,j}^k)^2.
\end{aligned} \tag{34}$$

Step III: Relationship for $F(\bar{x}^k)-F(\theta^*)$.

Because $F(\cdot)$ is convex with Lipschitz continuous gradients, we always have the following relation (see Theorem 2.1.5 in [45]):

$$F(u)\leq F(v)+\langle\nabla F(v),u-v\rangle+\frac{L}{2}\|v-u\|^2$$

for any $u,v\in\mathbb{R}^d$.

Letting $u=\bar{x}^{k+1}$ and $v=\bar{x}^k$ in the preceding relation yields

$$\begin{aligned}
&F(\bar{x}^{k+1}) \\
&\leq F(\bar{x}^k)+\langle\nabla F(\bar{x}^k),\bar{x}^{k+1}-\bar{x}^k\rangle+\frac{L}{2}\|\bar{x}^{k+1}-\bar{x}^k\|^2 \\
&\leq F(\bar{x}^k)+\langle\nabla F(\bar{x}^k),\gamma^k\bar{\zeta}_w^k-\gamma^k\bar{\xi}_w^k-\lambda^k\bar{g}^k\rangle \\
&\quad+\frac{L}{2}\|\gamma^k\bar{\zeta}_w^k-\gamma^k\bar{\xi}_w^k-\lambda^k\bar{g}^k\|^2,
\end{aligned} \tag{35}$$

where in the second inequality we used the relation in (27).

Subtracting $F(\theta^*)$ on both sides of (35) and then taking the expectation (conditioned on $\mathcal{F}^k$) on both sides yield

$$\begin{aligned}
&\mathbb{E}\left[F(\bar{x}^{k+1})-F(\theta^*)|\mathcal{F}^k\right] \\
&\leq F(\bar{x}^k)-F(\theta^*)-\langle\nabla F(\bar{x}^k),\lambda^k\bar{g}^k\rangle \\
&\quad+\frac{L}{2}\mathbb{E}\left[\|\gamma^k\bar{\zeta}_w^k-\gamma^k\bar{\xi}_w^k-\lambda^k\bar{g}^k\|^2\right] \\
&\leq F(\bar{x}^k)-F(\theta^*)-\langle\nabla F(\bar{x}^k),\lambda^k\bar{g}^k\rangle \\
&\quad+\frac{3L}{2}(\lambda^k)^2\|\bar{g}^k\|^2+\frac{3L}{2}(\gamma^k)^2\mathbb{E}\left[\|\bar{\zeta}_w^k\|^2\right] \\
&\quad+\frac{3L}{2}(\gamma^k)^2\mathbb{E}\left[\|\bar{\xi}_w^k\|^2\right] \\
&\leq F(\bar{x}^k)-F(\theta^*)-\langle\nabla F(\bar{x}^k),\lambda^k\bar{g}^k\rangle \\
&\quad+\frac{3L}{2}(\lambda^k)^2\|\bar{g}^k\|^2+\frac{3L}{2}(\gamma^k)^2\sum_{i,j}(R_{ij}\sigma_{\zeta,j}^k)^2 \\
&\quad+\frac{3L}{2}(\gamma^k)^2\sum_{i,j}(C_{ij}\sigma_{\xi,j}^k)^2.
\end{aligned} \tag{36}$$

Next we bound the inner product term. Using the relationship $-\langle a,b\rangle=\frac{\|a-b\|^2-\|a\|^2-\|b\|^2}{2}$ valid for any vectors a and b, one obtains

$$\begin{aligned}
&-\langle\nabla F(\bar{x}^k),\lambda^k\bar{g}^k\rangle \\
&=\frac{\lambda^k}{2}\left(\|\nabla F(\bar{x}^k)-\bar{g}^k\|^2-\|\nabla F(\bar{x}^k)\|^2-\|\bar{g}^k\|^2\right) \\
&\leq\frac{\lambda^k}{2}\left(\left\|\frac{1}{m}\sum_{i=1}^m(\nabla f_i(\bar{x}^k)-\nabla f_i(x_i^k))\right\|^2-\|\nabla F(\bar{x}^k)\|^2\right. \\
&\quad\left.-\|\bar{g}^k\|^2\right) \\
&\leq\frac{\lambda^k L^2}{2m}\sum_{i=1}^m\|x_i^k-\bar{x}^k\|^2-\frac{\lambda^k}{2}\|\nabla F(\bar{x}^k)\|^2-\frac{\lambda^k}{2}\|\bar{g}^k\|^2 \\
&=\frac{\lambda^k L^2}{2m}\|\mathbf{x}^k-\mathbf{1}\bar{x}^k\|_2^2-\frac{\lambda^k}{2}\|\nabla F(\bar{x}^k)\|^2-\frac{\lambda^k}{2}\|\bar{g}^k\|^2.
\end{aligned} \tag{37}$$

Plugging (37) into (36) leads to

$$\begin{aligned}
&\mathbb{E}\left[F(\bar{x}^{k+1})-F(\theta^*)|\mathcal{F}^k\right] \\
&\leq F(\bar{x}^k)-F(\theta^*)+\frac{\lambda^k L^2}{2m}\|\mathbf{x}^k-\mathbf{1}\bar{x}^k\|_2^2-\frac{\lambda^k}{2}\|\nabla F(\bar{x}^k)\|^2 \\
&\quad-\left(\frac{\lambda^k-3L(\lambda^k)^2}{2}\right)\|\bar{g}^k\|^2+\frac{3L}{2}(\gamma^k)^2\sum_{i,j}(R_{ij}\sigma_{\zeta,j}^k)^2 \\
&\quad+\frac{3L}{2}(\gamma^k)^2\sum_{i,j}(C_{ij}\sigma_{\xi,j}^k)^2.
\end{aligned} \tag{38}$$

Step IV: We combine Steps I-III and prove the theorem. Defining

$$\mathbf{v}^k=\left[F(\bar{x}^{k+1})-F(\theta^*),\|\mathbf{x}^k-\mathbf{1}\bar{x}^k\|_R^2,\|\mathbf{s}^k-v\bar{s}^k\|_C^2\right]^T,$$

we have the following relations from (26), (34), and (38):

$$\mathbb{E}\left[\mathbf{v}^{k+1}|\mathcal{F}^k\right]\leq(V^k+A^k)\mathbf{v}^k-H^k\begin{bmatrix}\|\nabla F(\bar{x}^k)\|^2\\\|\bar{g}^k\|^2\end{bmatrix}+B^k, \tag{39}$$

where

$$V^k = \begin{bmatrix} 1 & \frac{\delta_{2,R}^2 \lambda^k L^2}{2m} & 0 \\ 0 & 1-\gamma_1^k \rho_R & \frac{2\gamma^k \|\Pi_U C\|_R^2 \delta_{R,C}^2}{\rho_R} \\ 0 & 0 & 1-\gamma_2^k \rho_c \end{bmatrix},$$

$$A^k = \begin{bmatrix} 0 & 0 & 0 \\ a_1^k & a_2^k & 0 \\ a_3^k & a_4^k & 0 \end{bmatrix},$$

$$H^k = \begin{bmatrix} \frac{\lambda^k}{2} & \frac{\lambda^k - 3L(\lambda^k)^2}{2} \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \; B^k = \begin{bmatrix} b_1^k \\ b_2^k \\ b_3^k \end{bmatrix},$$

with

$$a_1^k = \frac{16mL(\lambda^k)^2 \|\Pi_U\|_R^2 \delta_{R,2}^2}{\gamma^k \rho_R},$$

$$a_2^k = \frac{4(\lambda^k)^2 L^2 \|\Pi_U\|_R^2 \delta_{R,2}^2}{\gamma^k \rho_R},$$

$$a_3^k = \frac{8mL(\lambda^k)^2 \delta_{C,2}^2 \|\Pi_v\|_C^2}{\gamma^k \rho_c},$$

$$a_4^k = \frac{2L^2(\lambda^k)^2 \delta_{C,2}^2 \delta_{2,R}^2 \|\Pi_v\|_C^2}{\gamma^k \rho_c},$$

$$b_1^k = \frac{3L}{2}(\gamma^k)^2 \sum_{i,j} (R_{ij}\sigma_{\zeta,j}^k)^2 + \frac{3L}{2}(\gamma^k)^2 \sum_{i,j} (C_{ij}\sigma_{\xi,j}^k)^2,$$

$$b_2^k = \frac{8(\lambda^k)^2 \|\Pi_U\|_R^2 \delta_{R,2}^2}{\gamma^k \rho_R} \sum_{i=1}^m \|\nabla f_i(\theta^*)\|^2 + 2(\gamma^k)^2 \|\Pi_u\|_R^2 \delta_{R,2}^2 \sum_{i,j} (R_{ij}\sigma_{\zeta,j}^k)^2 + 2(\gamma^k)^2 \|\Pi_U\|_R^2 \delta_{R,2}^2 \sum_{i,j} (C_{ij}\sigma_{\xi,j}^k)^2,$$

$$b_3^k = \frac{4(\lambda^k)^2 \delta_{C,2}^2 \|\Pi_v\|_C^2}{\gamma^k \rho_c} \sum_{i=1}^m \|\nabla f_i(\theta^*)\|^2 + (\gamma^k)^2 \delta_{C,2}^2 \|\Pi_v\|_C^2 \sum_{i,j} (C_{ij}\sigma_{\xi,j}^k)^2.$$

Under Assumption 3, and the conditions that $(\gamma^k)^2 \frac{(\lambda^k)^2}{\gamma^k}$ are summable in the theorem statement, it follows that all entries of the matrix B^k are summable almost surely. By defining $\hat{b}^k$ as the maximum element of B^k, we have $B^k \leq \hat{b}^k \mathbf{1}$. Therefore, $\mathbb{E}\left[F(\bar{x}^k) - F(\theta^*)|\mathcal{F}^k\right]$, $\mathbb{E}\left[\|\mathbf{x}^k - \mathbf{1}\bar{x}^k\|_R^2|\mathcal{F}^k\right]$, and $\mathbb{E}\left[\|\mathbf{s}^k - v\bar{s}^k\|_C^2|\mathcal{F}^k\right]$ for the iterates generated by the proposed algorithm satisfy the conditions of Theorem 1 and, hence, the results of Theorem 1 hold. ■

Remark 6. *The requirement on the decaying-factor γ^k and stepsize λ^k in the statement of Theorem 2 can be satisfied, for example, by setting $\gamma^k = \mathcal{O}(\frac{1}{k^a})$ and $\lambda^k = \mathcal{O}(\frac{1}{k^b})$ with $a, b \in \mathbb{R}$ satisfying $0.5 < a < b \leq 1$ and $2b - a > 1$. For example, setting $\gamma^k = \frac{c_1}{1+c_2 k^\iota}$ and $\lambda^k = \frac{c_3}{1+c_4 k}$ will satisfy the conditions for any exponent $0.5 < \iota < 1$, and positive coefficients c_1, c_2, c_3, and c_4.*

Remark 7. *Using the relationship in (9) and the definitions of $\bar{s}^k$, $\bar{\xi}^k$, and $\bar{g}^k$ in (10) and (11), one can obtain*

$$\bar{s}^{k+1} - \bar{s}^k = \gamma^k \bar{\xi}_w^k + \lambda^k \bar{g}^k,$$

i.e., $\bar{s}^{k+1} - \bar{s}^k$ tracks the global gradient. Combined with the proven result in Theorem 2 that all s_i^k converge to each other, and hence to the average $\bar{s}^k$ of all s_i^k, one can deduce that $s_i^{k+1} - s_i^k$ in (4) of the proposed Algorithm 1 indeed tracks the global gradient.

V. Online Estimation of the left eigenvector

In Algorithm 1, when the communication graph $\mathcal{G}_R$ is not balanced, a preprocessing approach can be used to estimate the left eigenvector u^T. In this section, inspired by the online eigenvector estimation algorithm in [46], we propose Algorithm 2 below, which allows individual agents to estimate the left eigenvector u^T locally on the fly while updating their optimization iterations in a distributed manner:

Algorithm 2: Robust gradient-tracking based distributed optimization with eigenvector estimation

Parameters: Stepsize λ^k and a decaying factor γ^k to suppress information-sharing noise;

Every agent i maintains two states x_i^k and s_i^k, which are initialized randomly with $x_i^0 \in \mathbb{R}^d$ and $s_i^0 \in \mathbb{R}^d$. Every agent i also maintains an eigenvector-estimation parameter $z_i^k \in \mathbb{R}^m$ initialized with $z_i^0 = \mathbf{e}_i \in \mathbb{R}^m$ where $\mathbf{e}_i$ has the ith element equal to one and all other elements equal to zero.

for $k = 1, 2, \cdots$ **do**

a) Agent i pushes s_i^k to each agent $l \in \mathbb{N}_{C,i}^{\text{out}}$, which will be received as $s_i^k + \xi_i^k$ due to information-sharing noise. And agent i pulls x_j^k from each $j \in \mathbb{N}_{R,i}^{\text{in}}$, which will be received as $x_j^k + \zeta_j^k$ due to information-sharing noise. Here the subscript R or C in neighbor sets indicates the neighbors with respect to the graphs induced by these matrices. Agent i also pulls z_j^k from each $j \in \mathbb{N}_{R,i}^{\text{in}}$.

b) Agent i chooses $\gamma^k > 0$ satisfying $1 + \gamma^k R_{ii} > 0$ and $1 + \gamma^k C_{ii} > 0$ with R_{ii} and C_{ii} defined in (3). Then, agent i updates its states as follows:

$$\begin{aligned} s_i^{k+1} =& (1+\gamma^k C_{ii}) s_i^k + \gamma^k \sum_{j \in \mathbb{N}_{C,i}^{\text{in}}} C_{ij}(s_j^k + \xi_j^k) \\ &+ \lambda^k \nabla f_i(x_i^k), \\ x_i^{k+1} =& (1+\gamma^k R_{ii}) x_i^k + \gamma^k \sum_{j \in \mathbb{N}_{R,i}^{\text{in}}} R_{ij}(x_j^k + \zeta_j^k) \\ &- \frac{s_i^{k+1} - s_i^k}{m z_{ii}^k}, \\ z_i^{k+1} =& z_i^k + \sum_{j \in \mathbb{N}_{R,i}^{\text{in}}} R_{ij}(z_j^k - z_i^k), \end{aligned} \tag{40}$$

where z_{ii}^k denotes the ith element of z_i^k.

c) **end**

In Algorithm 2, every agent uses the third update in (40) to locally estimate the left eigenvector of $I + \gamma^k R$. (Note that as discussed in Remark 3, the left eigenvector of $I + \gamma^k R$ is time-invariant and independent of γ^k. Also note that the update obtains an estimated eigenvector with row sum equal to one, and thus we scale the estimate by m to obtain u^T whose

row sum is required to be m.) Therefore, every agent i can use its local estimate z_i^k of the left eigenvector, which avoids using global information of u^T in Algorithm 1. It is worth noting that since z_i^k does not contain sensitive information, there is no need to add information-sharing noise to them to enable differential privacy. Moreover, the dimension of z_i^k is equal to the size of the network m. Thus, even in the case where the communication channel is noisy or coarse quantization is used for s-iterates and x-iterates, special effort (e.g., error-correction coding [47] or high-precision quantization) can be exploited to ensure that shared z_i^k messages are not contaminated by noises. Note that such special effort may not be feasible for the sharing of optimization variables (s-iterates and x-iterates) since the dimension of optimization variables can scale up to hundreds of millions in deep learning applications [22], which makes the cost for error-correction coding or high-precision quantization prohibitively high.

Next, we prove that Algorithm 2 can still ensure almost sure convergence of all agents to an optimal solution. To this end, we first characterize the estimation error of the eigenvector estimator:

Lemma 4. *Under Assumption 2, the iterates z_i^k in (40), after scaled by m, converge to the left eigenvector $u^T = [u_1, u_2, \cdots, u_m]^T$ of $I + \gamma^k R$ with a geometric rate, i.e., there exist $C > 0$ and $p \in (0, 1)$ satisfying the following inequality for any $i \in [m]$ and $k \geq 0$:*

$$\left| \frac{1}{m z_{ii}^k} - \frac{1}{u_i} \right| \leq C p^k, \tag{41}$$

where z_{ii}^k denotes the ith element of z_i^k.

Proof. From [46], we know that there exist $C_1 > 0$ and $p \in (0, 1)$ such that $\left| m z_{ii}^k - u_i \right| \leq C_1 p^k$ holds under the given conditions. According to [46], we also know that u_i and z_{ii}^k are strictly positive numbers. Therefore, using the relation $\left| \frac{1}{m z_{ii}^k} - \frac{1}{u_i} \right| = \frac{\left| m z_{ii}^k - u_i \right|}{m z_{ii}^k u_i}$, we know that there exist $C > 0$ such that (41) holds. ∎

Based on Lemma 4, we can prove the almost sure convergence of all agents to an optimal solution following the line of reasoning of Theorem 2:

Theorem 3. *Let Assumption 1, Assumption 2, and Assumption 3 hold. If $\{\gamma^k\}$ and $\{\lambda^k\}$ satisfy $\sum_{k=0}^{\infty} \gamma^k = \infty$, $\sum_{k=0}^{\infty} (\gamma^k)^2 < \infty$, $\sum_{k=0}^{\infty} \lambda^k = \infty$, $\sum_{k=0}^{\infty} \frac{(\lambda^k)^2}{\gamma^k} < \infty$, and $\lim_{k\to\infty} \lambda^k / \gamma^k = 0$, then the results of Theorem 1 hold for Algorithm 2.*

Proof. The proof follows the derivation of Theorem 2. Since the eigenvalue estimation process does not affect the dynamics of s_i^k, the relation for $\|\mathbf{s}^k - v\bar{s}^k\|_C$ in Step I of Theorem 2 still holds for Algorithm 2. Thus, we only need to show that we can establish relations for $\|\mathbf{x}^k - \mathbf{1}\bar{x}^k\|_R$ and $\mathbb{E}\left[F(\bar{x}^k) - F(\theta^*) | \mathcal{F}^k\right]$ that are similar to those in Theorem 2.

Similar to (27), denoting U as $\mathrm{diag}(u_1, u_2, \cdots, u_m)$ and Z^k as $\mathrm{diag}(m z_{11}^k, m z_{22}^k, \cdots, m z_{mm}^k)$, with z_{ii}^k denoting the ith element of z_i^k, we can obtain the following relationship for the x-iterates in Algorithm 2:

$$\begin{aligned}
&\bar{x}^{k+1} \\
=&\frac{u^T}{m}\mathbf{x}^{k+1} = \frac{u^T}{m}(R^k\mathbf{x}^k + \gamma^k\boldsymbol{\zeta}_w^k - (Z^k)^{-1}(\mathbf{s}^{k+1} - \mathbf{s}^k)) \\
=&\frac{u^T}{m}\left(R^k\mathbf{x}^k + \gamma^k\boldsymbol{\zeta}_w^k - \left(U^{-1} + (Z^k)^{-1} - U^{-1}\right)(\mathbf{s}^{k+1} - \mathbf{s}^k)\right) \\
=&\bar{x}^k + \gamma^k\bar{\zeta}_w^k - \frac{\mathbf{1}^T}{m}(\mathbf{s}^{k+1} - \mathbf{s}^k) \\
&- \frac{u^T((Z^k)^{-1} - U^{-1})}{m}(\mathbf{s}^{k+1} - \mathbf{s}^k)).
\end{aligned} \tag{42}$$

Hence, following the line of reasoning in the proof of Theorem 2, we can obtain

$$\begin{aligned}
\mathbf{x}^{k+1} - \mathbf{1}\bar{x}^{k+1} =& \bar{R}^k(\mathbf{x}^k - \mathbf{1}\bar{x}^k) - \gamma^k\Pi_U C(\mathbf{s}^k - v\bar{s}^k) \\
&- \gamma^k\Pi_U\boldsymbol{\xi}_w^k - \lambda^k\Pi_U\mathbf{g}^k + \gamma^k\Pi_u\boldsymbol{\zeta}_w^k \\
&+ \gamma^k\Pi_U^e(\mathbf{s}^k - v\bar{s}^k) + \gamma^k\Pi_U^e\boldsymbol{\xi}_w^k + \lambda^k\Pi_U^e\mathbf{g}^k,
\end{aligned} \tag{43}$$

where $\Pi_U^e = (I - \frac{\mathbf{1}u^T}{m})((Z^k)^{-1} - U^{-1})$ and we have used the relationship in (29) in the last equality.

In (43), the last three terms on the right hand side correspond to the influence of introducing the eigenvector estimator. Given that the elements of $(Z^k)^{-1} - U^{-1}$ diminish with a geometric rate according to Lemma 4, we deduce that the coefficient sequences for these terms are all summable, and hence they will only introduce terms with summable coefficients in the relationship for $\|\mathbf{x}^k - \mathbf{1}\bar{x}^k\|_R$, which will not affect the almost sure convergence. The same reasoning applies to the dynamics of $\bar{x}^{k+1} - \bar{x}^k$. More specifically, compared with Algorithm 1, Algorithm 2's eigenvector estimator (the last item on the right hand side of (42)) introduces three extra terms $\gamma^k\frac{u^T((Z^k)^{-1} - U^{-1})}{m}C(\mathbf{s}^k - v\bar{s}^k)$, $\gamma^k\frac{u^T((Z^k)^{-1} - U^{-1})}{m}\boldsymbol{\xi}_w^k$, and $\lambda^k\frac{u^T((Z^k)^{-1} - U^{-1})}{m}\mathbf{g}^k$ according to (29). From Lemma 4, we know that their coefficients all decrease with a geometric rate and hence are all summable. Therefore, these three extra terms only introduce items that have summable coefficient sequences in the relationship for $F(\bar{x}^k) - F(\theta^*)$, which will not affect the almost sure convergence.

In summary, we have that introducing the eigenvector estimator adds terms with summable coefficient sequences in the final inequality in (39), and hence will not affect the almost sure convergence results in Theorem 2. Therefore, we can still prove that the iterates generated by Algorithm 2 satisfy the conditions of Theorem 1 and, hence, the results of Theorem 1 hold for Algorithm 2. ∎

VI. Extension to Distributed Stochastic Gradient Methods

In many distributed optimization applications, individual agents do not have access to the precise gradient and hence have to use *noisy* local gradients for optimization. For example, in modern machine learning on massive datasets, evaluating the precise gradient using all available data can be extremely expensive in computation or even practically infeasible. So individual agents usually only compute inexact

estimates of the true gradients using a portion of the data points available to them [24]. Furthermore, in the era of Internet of Things, which connect massive low-cost sensing and communication devices, the data fed to optimization computations are usually subject to measurement noises [48]. In this section, we prove that the proposed algorithm can ensure all agents' almost sure convergence to an optimal solution even when the gradients are noisy.

As in most existing results on stochastic gradient methods, we make the following standard assumption on the stochasticity of individual agents' local gradients:

Assumption 4. *Every individual agent's local gradient g_i^k is an unbiased estimate of the true gradient $\nabla f_i(x_i^k)$ and has bounded variance, i.e.,*

$$\mathbb{E}\left[g_i^k\right] = \nabla f_i(x_i^k),\ \forall i$$
$$\mathbb{E}\left[\|g_i^k - \nabla f_i(x)\|^2\right] \leq \sigma^2,\ \forall i, x$$

where σ is some positive constant.

Theorem 4. *Let Assumptions 1-4 hold. If $\{\gamma^k\}$ and $\{\lambda^k\}$ satisfy $\sum_{k=0}^{\infty}\gamma^k = \infty$, $\sum_{k=0}^{\infty}(\gamma^k)^2 < \infty$, $\sum_{k=0}^{\infty}\lambda^k = \infty$, $\sum_{k=0}^{\infty}\frac{(\lambda^k)^2}{\gamma^k} < \infty$, and $\lim_{k\to\infty}\lambda^k/\gamma^k = 0$, then the results of Theorem 1 hold for the proposed Algorithm 1 and Algorithm 2 even when individual agents have access to only stochastic estimates of their true gradients.*

Proof. We use Algorithm 1 as an example to prove the results. Similar derivations apply to Algorithm 2 as well.

The goal is still to establish the relationship in (14), with the σ-field $\mathcal{F}^k = \{x_i^\ell, s_i^\ell; 0 \leq \ell \leq k,\ i \in [m]\}$. To this end, we organize the derivations into four steps: in Step I, Step II, and Step III, we establish respectively relations for $\|\mathbf{s}^k - v\bar{s}^k\|_C$, $\|\mathbf{x}^k - \mathbf{1}\bar{x}^k\|_R$, and $\mathbb{E}\left[F(\bar{x}^k) - F(\theta^*)|\mathcal{F}^k\right]$ for the iterates generated by the proposed algorithm. In Step IV, we use them to show that (14) of Theorem 1 holds.

Step I: Relationship for $\|\mathbf{s}^k - v\bar{s}^k\|_C$.

Since the noise on gradients can be grouped into the noise term ξ_i^k, following the same procedure as in Theorem 2, we can obtain a relation similar to (26):

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{s}^{k+1} - v\bar{s}^{k+1}\right\|_C^2|\mathcal{F}^k\right] &\leq (1-\gamma^k\rho_c)\left\|\mathbf{s}^k - v\bar{s}^k\right\|_C^2\\ &+ \frac{2L^2(\lambda^k)^2\delta_{C,2}^2\|\Pi_v\|_C^2}{\gamma^k\rho_c}\|\mathbf{x}^k - \mathbf{1}\bar{x}^k\|_2^2\\ &+ \frac{8mL(\lambda^k)^2\delta_{C,2}^2\|\Pi_v\|_C^2}{\gamma^k\rho_c}(F(\bar{x}^k) - F(\theta^*))\\ &+ \frac{4(\lambda^k)^2\delta_{C,2}^2\|\Pi_v\|_C^2}{\gamma^k\rho_c}\sum_{i=1}^m\|\nabla f_i(\theta^*)\|^2\\ &+ (\gamma^k)^2\delta_{C,2}^2\|\Pi_v\|_C^2\sum_{i,j}(C_{ij}\sigma_{\xi,j}^k)^2\\ &+ m^2(\lambda^k)^2\delta_{C,2}^2\|\Pi_v\|_C^2\sigma^2,\end{aligned} \tag{44}$$

where the last term corresponds to the influence caused by the stochasticity in local gradients.

Step II: Relationship for $\|\mathbf{x}^k - \mathbf{1}\bar{x}^k\|_R$.

Still following the derivations in Theorem 2, we have that the stochasticity in local gradients will affect the term $\|\mathbf{g}^k\|_R^2$ in (32). More specifically, after taking conditional expectation, $\|\mathbf{g}^k\|_R^2$ will become $\|\mathbf{g}^k\|_R^2 + m^2\delta_{R,2}^2\sigma^2$, and hence (34) becomes

$$\begin{aligned}&\mathbb{E}\left[\|\mathbf{x}^{k+1} - \mathbf{1}\bar{x}^{k+1}\|_R^2|\mathcal{F}^k\right]\\ &\leq \left(1 - \gamma^k\rho_R + \frac{4(\lambda^k)^2L^2\|\Pi_U\|_R^2\delta_{R,2}^2}{\gamma^k\rho_R}\right)\|\mathbf{x}^k - \mathbf{1}\bar{x}^k\|_R^2\\ &+ \frac{2\gamma^k\|\Pi_U C\|_R^2}{\rho_R}\|\mathbf{s}^k - v\bar{s}^k\|_R^2\\ &+ \frac{16mL(\lambda^k)^2\|\Pi_U\|_R^2\delta_{R,2}^2}{\gamma^k\rho_R}(F(\bar{x}^k) - F(\theta^*))\\ &+ \frac{8(\lambda^k)^2\|\Pi_U\|_R^2\delta_{R,2}^2}{\gamma^k\rho_R}\sum_{i=1}^m\|\nabla f_i(\theta^*)\|^2\\ &+ \frac{2m^2(\lambda^k)^2\|\Pi_U\|_R^2\delta_{R,2}^2}{\gamma^k\rho_R}\sigma^2\\ &+ 2(\gamma^k)^2\|\Pi_u\|_R^2\delta_{R,2}^2\sum_{i,j}(R_{ij}\sigma_{\zeta,j}^k)^2\\ &+ 2(\gamma^k)^2\|\Pi_U\|_R^2\delta_{R,2}^2\sum_{i,j}(C_{ij}\sigma_{\xi,j}^k)^2.\end{aligned} \tag{45}$$

Step III: Relationship for $F(\bar{x}^k) - F(\theta^*)$.

Following the derivations in Theorem 2, we have that the stochasticity in local gradients affects $\bar{g}^k$, which will be subject to noise with variance σ^2. More specifically, we have that (38) becomes

$$\begin{aligned}&\mathbb{E}\left[F(\bar{x}^{k+1}) - F(\theta^*)|\mathcal{F}^k\right]\\ &\leq F(\bar{x}^k) - F(\theta^*) + \frac{\lambda^k L^2}{2m}\|\mathbf{x}^k - \mathbf{1}\bar{x}^k\|_2^2 - \frac{\lambda^k}{2}\|\nabla F(\bar{x}^k)\|^2\\ &- \left(\frac{\lambda^k - 3L(\lambda^k)^2}{2}\right)\|\bar{g}^k\|^2 + \frac{3L}{2}(\lambda^k)^2\sigma^2\\ &+ \frac{3L}{2}(\gamma^k)^2\sum_{i,j}(R_{ij}\sigma_{\zeta,j}^k)^2 + \frac{3L}{2}(\gamma^k)^2\sum_{i,j}(C_{ij}\sigma_{\xi,j}^k)^2.\end{aligned} \tag{46}$$

Step IV: We combine Steps I-III and prove the theorem, which involves arguments exactly the same as in the proof of Theorem 2. In fact, following the derivation in Theorem 2, we can obtain that the stochasticity of gradients will only affect the matrix B^k in (39), which will still be summable. Therefore, we can arrive at the same conclusion as in Theorem 2 even when local gradients are stochastic.

■

VII. Numerical Simulations

In this section, we evaluate the performance of the proposed distributed optimization algorithm within the context of a distributed estimation problem.

We consider a canonical distributed estimation problem where a network of m sensors collectively estimate an unknown parameter $\theta \in \mathbb{R}^d$. More specifically, we assume that each sensor i has a noisy measurement of the parameter, $z_i = M_i\theta + w_i$, where $M_i \in \mathbb{R}^{s\times d}$ is the measurement matrix of agent i and w_i is Gaussian measurement noise of unit variance. Then the maximum likelihood estimation of parameter θ can be solved using the optimization problem formulated as

(1), with each $f_i(\theta)$ given as $f_i(\theta) = \|z_i - M_i\theta\|^2 + \varsigma\|\theta\|^2$, where ς is a regularization parameter [9].

In the numerical experiments, we set the number of agents (sensors) to $m = 100$ and adopt a random interaction graph. To ensure that the random interaction graph is strongly connected, we first arrange the 100 agents on a ring and then add a directed link between any two nonadjacent nodes with probability 0.3. In the evaluation, we set $s = 3$ and $d = 2$. To evaluate the performance of the proposed algorithms, we inject Gaussian based information-sharing noise ζ_i^k and ξ_i^k on all shared x_i^k and s_i^k, respectively. Both ζ_i^k and ξ_i^k have mean 0 and standard deviation $\sigma_{\zeta,i}^k = \sigma_{\xi,i}^k = 0.8$. We set the stepsize λ^k and diminishing sequence γ^k as $\lambda^k = \frac{0.02}{1+0.1k}$ and $\gamma^k = \frac{1}{1+0.1k^{0.6}}$, respectively, which satisfy the conditions in Theorem 2. We run our Algorithm 1 and Algorithm 2 for 100 times and calculate the average as well as the variance of the optimization error $\sum_{i=1}^m \|x_i^k - \theta^*\|$ as a function of the iteration index k. The result for Algorithm 1 is given by the black curve and error bars in Fig. 1, and the result for Algorithm 2 is given by the cyan curve and error bars in Fig. 1. For comparison, we also run the conventional Push-Pull algorithm in [18] (which uses a constant stepsize and no decaying factor), the robust gradient-tracking algorithm proposed in [38] (which uses a constant stepsize and can avoid noise accumulation under constant inter-agent coupling), and our recent result in [39] (which combines the conventional Push-Pull with decaying factors). The stepsize for the conventional Push-Pull method in [18] and the algorithm in [38] is set to a constant value $\lambda^k = 0.02$, and the decaying factor and stepsize for [39] is set the same as ours (note that [39] uses two decaying factors, and we set one of them equal to our decaying factor and the other one is selected according to the requirement therein). For all these three algorithms, we run the experiments for 100 times under the same information-sharing noise. The evolution of the average optimization errors and variances for the three algorithms are depicted by the curves and error bars in orange, magenta, and blue, respectively, in Fig. 1. It is clear that the proposed algorithms have both faster convergence speeds and better optimization accuracies compared with existing results. Furthermore, it can be seen that the variance of the optimization error for the conventional Push-Pull algorithm in [18] indeed grows with time, which corroborates the accumulation of information-sharing noise in conventional gradient-tracking based algorithms. It is also worth noting that for the approach in [38] with a constant stepsize, although the theoretical analysis therein establishes that the expected optimization error converges linearly to a steady-state value, the actual optimization error may decrease with a slower rate.

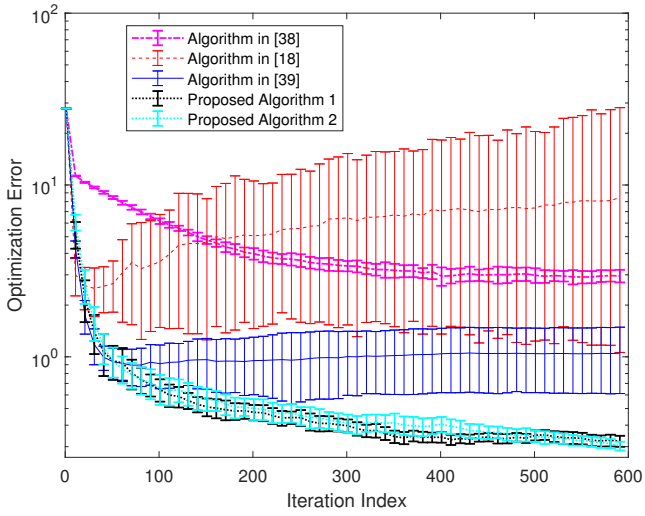


Fig. 1. Comparison of the proposed algorithms with the conventional Push-Pull algorithm in [18], the algorithm in [38] that can avoid noise accumulation when coupling matrices are constant, and the algorithm in [39] that combines conventional Push-Pull with decaying factors.

VIII. Conclusions

The robustness of distributed optimization algorithms against information-sharing noise is becoming increasingly important due to the prevalence of channel noise, the existence of quantization errors, and the demand for data perturbation/randomization for privacy protection. However, gradient-tracking based distributed optimization, which is gaining increased traction due to its applicability to general directed graphs and fast convergence speed, is vulnerable to information-sharing noise. In fact, in existing algorithms, information-sharing noise accumulates on the global gradient estimate and its variance will even grow to infinity when the noise is persistent. We have proposed a new gradient-tracking based approach which can avoid information-sharing noise from accumulating in the global-gradient estimate. The approach is applicable even when the inter-agent interaction is time-varying, enabling the incorporation of a decaying factor to gradually eliminate the influence of information-sharing noise, even when the noise is persistent. We have proved that with an appropriately chosen decaying factor, the proposed approach can guarantee all agents' almost sure convergence to an optimal solution for general convex objective functions with Lipschitz gradients, even in the presence of persistent information-sharing noise. The approach is also applicable when local gradients are subject to bounded noises as well, which is common in machine learning applications. Numerical simulation results confirm that in the presence of information-sharing noise, the proposed approach has better optimization accuracy compared with existing counterparts.

We should note that a limitation of our approach is that it assumes time-invariant coupling topology. We plan to explore relaxation of this assumption in future work. Moreover, in future work, we also plan to study whether decaying factors can be incorporated into non-gradient based distributed optimization algorithms to enable robustness against information-sharing noise.

References

[1] Tao Yang, Xinlei Yi, Junfeng Wu, Ye Yuan, Di Wu, Ziyang Meng, Yiguang Hong, Hong Wang, Zongli Lin, and Karl H Johansson. A survey of distributed optimization. *Annual Reviews in Control*, 47:278–305, 2019.

[2] Juan Andrés Bazerque and Georgios B Giannakis. Distributed spectrum sensing for cognitive radio networks by exploiting sparsity. *IEEE Transactions on Signal Processing*, 58(3):1847–1862, 2009.

[3] Robin L Raffard, Claire J Tomlin, and Stephen P Boyd. Distributed optimization for cooperative agents: Application to formation flight. In *2004 43rd IEEE Conference on Decision and Control*, pages 2453–2459, 2004.

[4] Chunlei Zhang and Yongqiang Wang. Distributed event localization via alternating direction method of multipliers. *IEEE Transactions on Mobile Computing*, 17(2):348–361, 2017.

[5] Konstantinos I Tsianos, Sean Lawlor, and Michael G Rabbat. Consensus-based distributed optimization: Practical issues and applications in large-scale machine learning. In *Proceedings of the 50th annual Allerton Conference on Communication, Control, and Computing*, pages 1543–1550, 2012.

[6] John N. Tsitsiklis. Problems in decentralized decision making and computation. Technical report, MIT, 1984.

[7] Angelia Nedic and Asuman Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.

[8] Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. EXTRA: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.

[9] Jinming Xu, Shanying Zhu, Yeng Chai Soh, and Lihua Xie. Convergence of asynchronous distributed gradient methods over stochastic networks. *IEEE Transactions on Automatic Control*, 63(2):434–448, 2017.

[10] Guannan Qu and Na Li. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260, 2017.

[11] Ran Xin and Usman A Khan. A linear algorithm for optimization over directed graphs with geometric convergence. *IEEE Control Systems Letters*, 2(3):315–320, 2018.

[12] Wei Shi, Qing Ling, Kun Yuan, Gang Wu, and Wotao Yin. On the linear convergence of the ADMM in decentralized consensus optimization. *IEEE Transactions on Signal Processing*, 62(7):1750–1761, 2014.

[13] Chunlei Zhang, Muaz Ahmad, and Yongqiang Wang. ADMM based privacy-preserving decentralized optimization. *IEEE Transactions on Information Forensics and Security*, 14(3):565–580, 2019.

[14] Ermin Wei, Asuman Ozdaglar, and Ali Jadbabaie. A distributed Newton method for network utility maximization–I: Algorithm. *IEEE Transactions on Automatic Control*, 58(9):2162–2175, 2013.

[15] Jiaqi Zhang, Keyou You, and Tamer Başar. Distributed adaptive Newton methods with global superlinear convergence. *Automatica*, 138:110156, 2022.

[16] Kun Yuan, Qing Ling, and Wotao Yin. On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3):1835–1854, 2016.

[17] Paolo Di Lorenzo and Gesualdo Scutari. NEXT: In-network nonconvex optimization. *IEEE Transactions on Signal and Information Processing over Networks*, 2(2):120–136, 2016.

[18] Shi Pu, Wei Shi, Jinming Xu, and Angelia Nedić. Push-pull gradient methods for distributed optimization in networks. *IEEE Transactions on Automatic Control*, 66(1):1–16, 2021.

[19] Soummya Kar and José MF Moura. Distributed consensus algorithms in sensor networks with imperfect communication: Link failures and channel noise. *IEEE Transactions on Signal Processing*, 57(1):355–369, 2008.

[20] Kunal Srivastava and Angelia Nedic. Distributed asynchronous constrained stochastic optimization. *IEEE Journal of Selected Topics in Signal Processing*, 5(4):772–790, 2011.

[21] Edward A Lee and David G Messerschmitt. *Digital Communication*. Springer Science & Business Media, 2012.

[22] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[23] Wei Wen, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Terngrad: ternary gradients to reduce communication in distributed deep learning. In *31st International Conference on Neural Information Processing Systems (NIPS 2017)*, 2017.

[24] Anastasia Koloskova, Tao Lin, Sebastian U Stich, and Martin Jaggi. Decentralized deep learning with arbitrary communication compression. In *International Conference on Learning Representations*, 2020.

[25] Xuanyu Cao and Tamer Başar. Decentralized online convex optimization with compressed communications. *submitted*.

[26] Chunlei Zhang and Yongqiang Wang. Enabling privacy-preservation in decentralized optimization. *IEEE Transactions on Control of Network Systems*, 6(2):679–689, 2018.

[27] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 691–706. IEEE, 2019.

[28] Yongqiang Wang and H Vincent Poor. Decentralized stochastic optimization with inherent privacy protection. *IEEE Transactions on Automatic Control*, 2022.

[29] Shuo Han, Ufuk Topcu, and George J Pappas. Differentially private distributed constrained optimization. *IEEE Transactions on Automatic Control*, 62(1):50–64, 2016.

[30] Yu Wang, Zhenqi Huang, Sayan Mitra, and Geir E Dullerud. Differential privacy in linear distributed control systems: Entropy minimizing mechanisms and performance tradeoffs. *IEEE Transactions on Control of Network Systems*, 4(1):118–130, 2017.

[31] Xueru Zhang, Mohammad Mahdi Khalili, and Mingyan Liu. Recycled admm: Improving the privacy and accuracy of distributed algorithms. *IEEE Transactions on Information Forensics and Security*, 15:1723–1734, 2019.

[32] Jianping He, Lin Cai, and Xinping Guan. Differential private noise adding mechanism and its application on consensus algorithm. *IEEE Transactions on Signal Processing*, 68:4069–4082, 2020.

[33] Jiaqi Zhang, Keyou You, and Tamer Başar. Distributed discrete-time optimization in multiagent networks using only sign of relative state. *IEEE Transactions on Automatic Control*, 64(6):2352–2367, 2018.

[34] Xuanyu Cao and Tamer Başar. Decentralized online convex optimization based on signs of relative states. *Automatica*, 129:109676, 2021.

[35] Jemin George, Tao Yang, He Bai, and Prudhvi Gurram. Distributed stochastic gradient method for non-convex problems with applications in supervised learning. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 5538–5543. IEEE, 2019.

[36] Thinh T Doan, Siva Theja Maguluri, and Justin Romberg. Convergence rates of distributed gradient methods under random quantization: A stochastic approximation approach. *IEEE Transactions on Automatic Control*, 66(10):4469–4484, 2021.

[37] Yongqiang Wang and Tamer Başar. Quantization enabled privacy protection in decentralized stochastic optimization. *IEEE Transactions on Automatic Control*, 2022.

[38] Shi Pu. A robust gradient tracking method for distributed optimization over directed networks. In *59th IEEE Conference on Decision and Control, extended version available at https://arxiv.org/pdf/2003.13980.pdf*, pages 2335–2341. IEEE, 2020.

[39] Yongqiang Wang and Angelia Nedić. Tailoring gradient methods for differentially-private distributed optimization. *http://arxiv.org/abs/2202.01113*, 2022.

[40] Fakhteh Saadatniaki, Ran Xin, and Usman A Khan. Decentralized optimization over time-varying directed graphs with row and column-stochastic matrices. *IEEE Transactions on Automatic Control*, 65(11):4769–4780, 2020.

[41] Shi Pu and Angelia Nedić. Distributed stochastic gradient tracking methods. *Mathematical Programming*, 187(1):409–457, 2021.

[42] Ran Xin, Anit Kumar Sahu, Usman A Khan, and Soummya Kar. Distributed stochastic optimization with gradient tracking over strongly-connected networks. In *IEEE Conference on Decision and Control*, pages 8353–8358, 2019.

[43] Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, and Ramtin Pedarsani. An exact quantized decentralized gradient descent algorithm. *IEEE Transactions on Signal Processing*, 67(19):4934–4947, 2019.

[44] Roger A Horn and Charles R Johnson. *Matrix Analysis*. Cambridge university press, 2012.

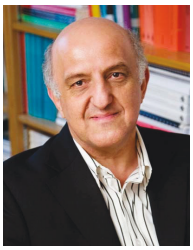
[45] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2003.

[46] Van Sy Mai and Eyad H Abed. Distributed optimization over weighted directed graphs using row stochastic matrix. In *American Control Conference (ACC)*, pages 7165–7170. IEEE, 2016.

[47] George C Clark Jr and J Bibb Cain. *Error-correction Coding for Digital Communications*. Springer Science & Business Media, 2013.

[48] Ran Xin, Soummya Kar, and Usman A Khan. Decentralized stochastic optimization and machine learning: A unified variance-reduction framework for robust performance and fast convergence. *IEEE Signal Processing Magazine*, 37(3):102–113, 2020.

Yongqiang Wang was born in Shandong, China. He received dual B.S. degrees in electrical engineering & automation and computer science & technology from Xi'an Jiaotong University, Xi'an, Shaanxi, China, in 2004, and the Ph.D. degree in control science and engineering from Tsinghua University, Beijing, China, in 2009. From 2007-2008, he was with the University of Duisburg-Essen, Germany, as a visiting student. He was a Project Scientist at the University of California, Santa Barbara before joining Clemson University, SC, USA, where he is currently an Associate Professor. His current research interests include distributed control, optimization, and learning, with an emphasis on privacy protection. He currently serves as an associate editor for *IEEE Transactions on Automatic Control* and *IEEE Transactions on Control of Network Systems*.



Tamer Başar (S'71-M'73-SM'79-F'83-LF'13) has been with the University of Illinois Urbana-Champaign since 1981, where he is currently Swanlund Endowed Chair Emeritus and Center for Advanced Study (CAS) Professor Emeritus of Electrical and Computer Engineering, with also affiliations with the Coordinated Science Laboratory, Information Trust Institute, and Mechanical Science and Engineering. At Illinois, he has also served as Director of CAS (2014-2020), Interim Dean of Engineering (2018), and Interim Director of the Beckman Institute (2008-2010). He received B.S.E.E. from Robert College, Istanbul, and M.S., M.Phil, and Ph.D. from Yale University, from which he received in 2021 the Wilbur Cross Medal. He is a member of the US National Academy of Engineering, and Fellow of IEEE, IFAC, and SIAM. He has served as presidents of IEEE CSS (Control Systems Society), ISDG (International Society of Dynamic Games), and AACC (American Automatic Control Council). He has received several awards and recognitions over the years, including the highest awards of IEEE CSS, IFAC, AACC, and ISDG, the IEEE Control Systems Award, and a number of international honorary doctorates and professorships. He has around 1000 publications in systems, control, communications, optimization, networks, and dynamic games, including books on non-cooperative dynamic game theory, robust control, network security, wireless and communication networks, and stochastic networked control. He was the Editor-in-Chief of Automatica between 2004 and 2014, and is currently editor of several book series. His current research interests include stochastic teams, games, and networks; multi-agent systems and learning; data-driven distributed optimization; epidemics modeling and control over networks; security and trust; energy systems; and cyber-physical systems.