Convergence of the Distributed SG Algorithm Under Cooperative Excitation Condition

Die Gan and Zhixin Liu, *Member, IEEE*

***Abstract*— In this article, a distributed stochastic gradient (SG) algorithm is proposed where the estimators are aimed to collectively estimate an unknown time-invariant parameter from a set of noisy measurements obtained by distributed sensors. The proposed distributed SG algorithm combines the consensus strategy of the estimation of neighbors with the diffusion of regression vectors. For the theoretical investigation of the proposed algorithm, the main challenge lies in analyzing the influence of the Laplacian matrix on the state transition matrix and the properties of the product of nonindependent and nonstationary random matrices. Some analysis techniques such as graph theory and martingale theory are used to deal with the above issues. A cooperative excitation condition is introduced, under which the convergence of the distributed SG algorithm can be obtained without relying on the independency or stationarity assumptions of regression vectors which are commonly used in the existing literature. Furthermore, the convergence rate of the algorithm can be established. Finally, we show that all the sensors can cooperate to fulfill the estimation task even though any individual sensor cannot by a simulation example.**

***Index Terms*— Convergence, cooperative excitation condition, distributed estimation, stochastic dynamic system, stochastic gradient (SG) algorithm.**

I. Introduction

PARAMETER estimation or filtering is one of the important issues in diverse fields including statistical learning, signal processing, system identification, and adaptive control. With the development of computer science and communication, sensor networks are widely applied due to the advantages of flexibility, fault tolerance, and ease of deployment. The sensor networks bring more and more data, and how to apply the data to design proper estimation algorithms is a promising research direction.

Manuscript received 12 November 2021; revised 14 May 2022 and 19 August 2022; accepted 4 October 2022. Date of publication 25 October 2022; date of current version 3 May 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFA0703800, in part by the Strategic Priority Research Program of Chinese Academy of Sciences under Grant XDA27000000, in part by the Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0103, and in part by the National Science Foundation of Shandong Province under Grant ZR2020ZD26. *(Corresponding author: Zhixin Liu.)*

Die Gan is with the Zhongguancun Laboratory, and the Key Laboratory of Systems and Control, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China (e-mail: gandie@zgclab.edu.cn).

Zhixin Liu is with the Key Laboratory of Systems and Control, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100190, China (e-mail: lzx@amss.ac.cn).

Digital Object Identifier 10.1109/TNNLS.2022.3213715

Generally speaking, there are three manners to process the information from the sensors: centralized, distributed, and a combination of both. For the centralized method, the information measured by the sensors is transmitted to a fusion center which uses the information to estimate the unknown signals or parameters. Compared with the distributed algorithms, the centralized ones lack robustness and bring a large amount of computation and communication burden. In distributed algorithms, the sensors can accomplish complicated tasks in a cooperative manner even though each sensor can only receive local information. A number of theoretical results on distributed estimation or learning algorithms (e.g., [1], [2], [3]) arise because of comprehensive practical applications in engineering systems, such as target localization and collaborative spectral sensing, see, e.g., [4], [5].

In the investigation of distributed estimation algorithms, how to use the local information to design the algorithms is important for the property of the algorithms. Three types of strategies are often adopted in the current literature: incremental strategy (cf., [6]), consensus strategy (cf., [7]), and diffusion strategy (cf., [8], [9]). Based on these three strategies, many different distributed adaptive estimation algorithms are proposed, such as the diffusion least mean squares (LMS), the consensus-based Kalman filter, and the diffusion least squares. Correspondingly, the stability and the convergence analysis of the distributed estimation algorithms are also investigated under some signal conditions (cf., [10], [11], [12], [13], [14], [15], [16], [17], [18], [19]). Schizas et al. [15] established the stability results for a distributed LMS-type adaptive algorithm with the strictly stationary ergodic regressor vectors. Takahashi et al. [16] investigated the mean transient and mean-square performance analysis of the diffusion LMS algorithm with the independent and identically distributed regressors. Cattivelli and Sayed [17] provided the steady-state mean and mean-square analysis of the diffusion LMS algorithm with the independent Gaussian regressors. Arablouei et al. [18] presented the convergence analysis of a partial-diffusion recursive least squares algorithm for the independent and ergodic input vectors. Lei and Chen [19] studied the convergence of the distributed stochastic approximation algorithm with ergodic signals. So far, most results require that the regression signals satisfy some stringent conditions, such as independency, stationarity, and ergodicity assumptions, which makes it hard or even impossible to apply these theoretical results to practical feedback control systems such as autoregressive moving-average with exogenous input (ARMAX) model and Hammerstein system where the regressors are often generated

by the past input and output signals. We remark that a preliminary attempt toward the relaxation of the independency and stationarity assumptions was made by Chen et al. [20], where they provided a cooperative excitation condition to guarantee the stability of the diffusion LMS algorithm, and some elegant results for the distributed LMS algorithm were further established by Xie and Guo [8] and [21] under a general cooperative information condition.

It is well-known that the standard stochastic gradient (SG) algorithm has the advantages of simple expression and easy computation. The SG algorithm is widely applied in the area of adaptive control and also has deep connections with the SG descent algorithm and its variants which are widely used to deal with optimization problems in machine learning [22]. With the development of sensor networks, the distributed implementations of SG descent algorithms have attracted much attention of researchers (cf., [23], [24], [25], [26], [27], [28], [29], [30], [31]).

In this article, we consider a network of sensors which are aimed to collectively estimate an unknown time-invariant parameter. We propose a distributed SG algorithm based on the combined diffusion and consensus strategies and study the convergence properties of the proposed algorithm for a dynamic system. The analysis of the influence of the Laplacian matrix corresponding to the communication graph on the state transition matrix and the analysis of properties of the product of nonindependent and nonstationary random matrices bring challenges to us. The main contributions of this article are summarized as follows.

1) We propose a novel distributed SG algorithm where each sensor is only allowed to communicate with its neighbors. The information of the regression vectors is first diffused through the sensor networks, and then the estimation of the unknown parameters is obtained using the consensus-based strategy.
2) By introducing a cooperative excitation condition on the regressor signals which is weaker than the persistent excitation (PE) condition commonly used in the literature (see e.g., [32], [33], [34]), the strong consistency of the distributed SG algorithm can be established. By the cooperative excitation condition, we see that the estimation task can be still fulfilled by the cooperation of multiple sensors even if any of them cannot. Our results can be degenerated to the convergence results on the standard SG algorithm (cf., [35], [36]).
3) We finally establish the convergence rate of the distributed SG algorithm under the cooperative excitation condition. Different from the convergence analysis of the distributed SG descent algorithms in most existing literature where the data are required to satisfy the independent and identically distributed (i.i.d.) condition (see e.g., [25], [26]), our theoretical results are obtained without relying on such stringent assumptions of the system signals, which makes it possible for applications to the stochastic feedback systems. We use the graph theory, martingale theory, and the specific structure of the proposed distributed SG algorithm to overcome the difficulties arising in the analysis of the product of nonindependent and nonstationary random matrices.

The rest of this article is organized as follows. In Section II, we first propose the distributed SG algorithm and introduce the cooperative excitation condition. A necessary and sufficient condition for the strong consistency of the proposed algorithm and the conditions of the regressors for the convergence of the algorithm are given in Section III. The convergence rate of the distributed SG algorithm is given in Section IV. A simulation example is given in Section V to illustrate our theoretical results. The concluding remarks are made in the last section.

II. Problem Formulation

A. Some Preliminaries

In this article, we use $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ to denote an $m \times n$-dimensional matrix. For a matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\|$ denotes its Euclidean norm, i.e., $\|\boldsymbol{A}\| \triangleq (\lambda_{\max}(\boldsymbol{A}\boldsymbol{A}^T))^{(1/2)}$, where the notation T denotes the transpose operator and $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue of the matrix. The notations $\det(\cdot)$ and $\mathrm{tr}(\cdot)$ are used to denote the determinant and trace of the corresponding matrix respectively. If all the elements of a matrix are nonnegative, then it is a nonnegative matrix, and furthermore, if $\sum_{j=1}^{n} a_{ij} = 1$ for all i, then it is called a stochastic matrix. The Kronecker product $\boldsymbol{A} \otimes \boldsymbol{B}$ of two matrices $\boldsymbol{A} = (a_{ij}) \in \mathbb{R}^{m\times n}$ and $\boldsymbol{B} \in \mathbb{R}^{p\times q}$ is defined as

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{pmatrix} a_{11}\boldsymbol{B} & \cdots & a_{1n}\boldsymbol{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\boldsymbol{B} & \cdots & a_{mn}\boldsymbol{B} \end{pmatrix} \in \mathbb{R}^{mp\times nq}.$$

Our purpose is to propose a distributed estimation algorithm based on the information from a set of sensors and investigate the convergence properties of the proposed algorithm. The sensors in sensor networks are modeled as nodes, and the relationship between sensors is presented as an undirected weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$, where $\mathcal{V} = \{1, 2, 3, \ldots, n\}$ is the set of sensors (i.e., nodes), the edge set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ denotes the communication between sensors, and $\mathcal{A} = (a_{ij})$ is the weighted matrix. The elements of the matrix $\mathcal{A}$ satisfy: $a_{ij} > 0$ if $(i, j) \in \mathcal{E}$ and $a_{ij} = 0$ otherwise. The neighbor set of the sensor i is denoted as $N^i = \{j \in \mathcal{V}, (i, j) \in \mathcal{E}\}$, and the sensor i is also included in this set. Each sensor can only exchange information with its neighbors. A path of length ℓ is a sequence of nodes $\{i_1, \ldots, i_\ell\}$ satisfying $(i_j, i_{j+1}) \in \mathcal{E}$ for all $1 \le j \le \ell - 1$. The graph $\mathcal{G}$ is called connected if for any two sensors i and j, there is a path connecting them. The diameter $D(\mathcal{G})$ of the graph $\mathcal{G}$ is defined as the maximum shortest length of paths between any two sensors.

For simplicity of analysis, the properties of the distributed algorithm are considered under the condition that the weighted matrix $\mathcal{A}$ is symmetric and stochastic. Hence, the Laplacian matrix $\boldsymbol{L}$ of the graph $\mathcal{G}$ can be written as $\boldsymbol{L} = \boldsymbol{I} - \mathcal{A}$ with $\boldsymbol{I}$ being the identity matrix. According to [37], we can obtain the following properties about the Laplacian matrix $\boldsymbol{L}$.

Lemma 1: The Laplacian matrix $\boldsymbol{L}$ defined above has at least one zero eigenvalue, with other eigenvalues positive and less than or equal to 2. Moreover, if the graph $\mathcal{G}$ is connected, then $\boldsymbol{L}$ has only one zero eigenvalue.

The following lemma is often used in our analysis and we list it as follows.

Lemma 2 [38]: Let $D_t \triangleq 1 + \sum_{j=1}^{t} d_j,\ d_j \geq 0$, then

$$\sum_{j=1}^{\infty} \frac{d_j}{D_j^{\alpha}} < \infty \quad \forall\ \alpha > 1$$

$$\sum_{j=1}^{\infty} \frac{d_j}{D_j} = \infty, \quad \text{iff} \quad \lim_{j\to\infty} D_j = \infty.$$

B. Distributed SG Algorithm

In this article, we consider a network consisting of n sensors. The signal model of each sensor $i \in \{1, \ldots, n\}$ is assumed to obey the following time-invariant regression stochastic model:

$$y_{k+1}^i = \boldsymbol{\theta}^T \boldsymbol{\varphi}_k^i + \varepsilon_{k+1}^i, \quad k \geq 0 \tag{1}$$

where y_k^i is the scalar observation of the sensor i at the time instant k, $\boldsymbol{\varphi}_k^i \in \mathbb{R}^m$ is the random regression vector which may be the function of the current and past inputs and outputs, $\boldsymbol{\theta} \in \mathbb{R}^m$ is an unknown parameter to be estimated, and $\{\varepsilon_k^i\}$ is a noise process.

We aim at designing a distributed adaptive estimation algorithm where all the sensors cooperatively estimate the unknown parameter $\boldsymbol{\theta}$ of the stochastic dynamical system (1) using local information $\{y_{k+1}^j, \boldsymbol{\varphi}_k^j\}_{j\in N^i}$, and further establishing the (almost sure) convergence property and the convergence rate of the proposed distributed algorithm.

The standard SG algorithm is commonly used in the area of adaptive control and system identification (e.g., [36], [39]). Inspired by Liu et al. [1] and based on the standard SG algorithm in the context of stochastic adaptive control, we propose the following distributed SG algorithm to cooperatively estimate the unknown parameter $\boldsymbol{\theta}$. The detailed algorithm can be found in Algorithm 1.

Remark 3: Algorithm 1 designed by combining the consensus strategy of the estimation of neighbors with the diffusion strategy of regression vectors is online and updated from time to time using new measurement data. We see that the right-hand side of (2) in Algorithm 1 consists of two parts: the first part is the standard SG algorithm which tries to minimize the prediction error using the innovation, while the second part can be regarded as the result of minimizing the weighted distance between the estimates of the sensor i and its neighbors.

Remark 4: In Step 2, the multistep diffusion strategy of regression vectors is used, which was widely used in the design of the distributed algorithms (e.g., [40], [41], [42]). The diffusion step Q plays an important role in establishing the contraction property of the product of random matrices $\prod_{p=j}^{k}(\boldsymbol{I}_{mn} - \mu \boldsymbol{G}_p)$ (see the proof of Theorem 25 in Appendix D). Moreover, by following the proof of Lemma 15, we see that if the condition number of $\boldsymbol{\Phi}_k^T \boldsymbol{\Phi}_k$ is bounded, then the multistep diffusion step can be removed.

For convenience of analysis, we introduce the following notations (see Table I). In Table I, $\mathrm{col}(\cdot, \ldots, \cdot)$ denotes the vector stacked by the specified vectors, and $\mathrm{diag}(\cdot, \ldots, \cdot)$ denotes the block matrix formed in a diagonal manner of the corresponding vectors or matrices.

Algorithm 1 Distributed SG Algorithm

Input: $\{\boldsymbol{\varphi}_k^i, y_{k+1}^i\}_{i=1}^n,\ k = 0, 1, 2, \ldots$

Output: $\{\hat{\boldsymbol{\theta}}_{k+1}^i\}_{i=1}^n,\ k = 0, 1, 2, \ldots$

Initialization: For every sensor $i \in \{1, \ldots, n\}$, begin with an arbitrary initial vector $\hat{\boldsymbol{\theta}}_0^i$.

for each time $k = 0, 1, 2, \ldots$ **do**

for every sensor $i = 1, \ldots, n$ **do**

Step 1. Set the value as

$$x_k^i(0) = \frac{\left\|\boldsymbol{\varphi}_k^i\right\|^2}{r_k^i}, \quad r_k^i \triangleq 1 + \sum_{j=1}^{k} \left\|\boldsymbol{\varphi}_j^i\right\|^2.$$

Step 2. Perform the following diffusion process for $q = 0, 1, 2, \ldots, Q$ with $Q \geq D(\mathscr{G})$:

$$x_k^i(q+1) = \sum_{j\in N^i} a_{ij} x_k^j(q).$$

Step 3. Update the estimate $\hat{\boldsymbol{\theta}}_{k+1}^i$ of the unknown parameter,

$$z_k^i(\hat{\boldsymbol{\theta}}_k^i) = x_k^i(Q) \sum_{l\in N^i} a_{li}(\hat{\boldsymbol{\theta}}_k^i - \hat{\boldsymbol{\theta}}_k^l)$$

$$\hat{\boldsymbol{\theta}}_{k+1}^i = \underbrace{\hat{\boldsymbol{\theta}}_k^i + \mu \frac{\boldsymbol{\varphi}_k^i}{r_k^i}\left(y_{k+1}^i - (\boldsymbol{\varphi}_k^i)^T \hat{\boldsymbol{\theta}}_k^i\right)}_{\text{Standard SG}} \underbrace{- \mu\nu \sum_{j\in N^i} a_{ij}\left(z_k^i(\hat{\boldsymbol{\theta}}_k^i) - z_k^j(\hat{\boldsymbol{\theta}}_k^j)\right)}_{\text{Consensus-based item}} \tag{2}$$

where μ and ν are two step sizes lying in $(0, 1)$, and r_k^i is defined in Step 1.

TABLE I
SOME NOTATIONS

Notation	Definition	Dimension
$\boldsymbol{Y}_k$	$\{y_k^1, \cdots, y_k^n\}$	$1 \times n$
$\boldsymbol{\Phi}_k$	$diag\{\boldsymbol{\varphi}_k^1, \cdots, \boldsymbol{\varphi}_k^n\}$	$mn \times n$
$\boldsymbol{\Xi}_k$	$\{\varepsilon_k^1, \cdots, \varepsilon_k^n\}$	$1 \times n$
$\boldsymbol{\Theta}$	$col\{\underbrace{\boldsymbol{\theta}, \cdots, \boldsymbol{\theta}}_{n}\}$	$mn \times 1$
$\hat{\boldsymbol{\Theta}}_k$	$col\{\hat{\boldsymbol{\theta}}_k^1, \cdots, \hat{\boldsymbol{\theta}}_k^n\}$	$mn \times 1$
$\tilde{\boldsymbol{\Theta}}_k$	$col\{\tilde{\boldsymbol{\theta}}_k^1, \cdots, \tilde{\boldsymbol{\theta}}_k^n\},\ \tilde{\boldsymbol{\theta}}_k^i = \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_k^i$	$mn \times 1$
$\boldsymbol{R}_k$	$diag\{r_k^1, \cdots, r_k^n\}$	$n \times n$
$\mathscr{L}$	$\boldsymbol{L} \otimes \boldsymbol{I}_m,\ \boldsymbol{L}$ is the Laplacian matrix	$mn \times mn$
$\boldsymbol{A}_k$	$\boldsymbol{\Phi}_k \boldsymbol{R}_k^{-1} \boldsymbol{\Phi}_k^T$	$mn \times mn$
$\boldsymbol{X}_k(Q)$	$diag\{x_k^1(Q), \cdots, x_k^n(Q)\}$	$n \times n$
$\boldsymbol{G}_k$	$\boldsymbol{A}_k + \nu\mathscr{L}(\boldsymbol{X}_k(Q) \otimes \boldsymbol{I}_m)\mathscr{L}$	$mn \times mn$

Using the notations in Table I, we can rewrite (1) and (2) into the following compact form:

$$\begin{aligned} \boldsymbol{Y}_{k+1} &= \boldsymbol{\Theta}^T \boldsymbol{\Phi}_k + \boldsymbol{\Xi}_{k+1} \\ \hat{\boldsymbol{\Theta}}_{k+1} &= \hat{\boldsymbol{\Theta}}_k + \mu \boldsymbol{\Phi}_k \boldsymbol{R}_k^{-1}\left(\boldsymbol{Y}_{k+1}^T - \boldsymbol{\Phi}_k^T \hat{\boldsymbol{\Theta}}_k\right) \\ &\quad - \mu\nu\mathscr{L}(\boldsymbol{X}_k(Q) \otimes \boldsymbol{I}_m)\mathscr{L}\hat{\boldsymbol{\Theta}}_k. \end{aligned} \tag{3}$$

From the definition of $\boldsymbol{R}_k$ and (3), the term $\boldsymbol{R}_k^{-1}$ can be regarded as an adaptive gain matrix.

Let $\widetilde{\boldsymbol{\Theta}}_k = \boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}}_k$. It is clear that $\mathscr{L}\boldsymbol{\Theta} = 0$, and then we have the following error equation:

$$\widetilde{\boldsymbol{\Theta}}_{k+1} = (\boldsymbol{I}_{mn} - \mu \boldsymbol{G}_k)\widetilde{\boldsymbol{\Theta}}_k - \mu \boldsymbol{\Phi}_k \boldsymbol{R}_k^{-1} \boldsymbol{\Xi}_{k+1}^T. \tag{4}$$

To proceed with our analysis, we introduce some assumptions concerning the graph, regression vectors, and system noise.

Assumption 5: The undirected graph $\mathscr{G}$ is connected.

Assumption 6 (Nonpersistent Cooperative Excitation Condition): There exist two positive constants N and K_0 such that for $k \geq K_0$, the following inequality is satisfied:

$$\frac{\lambda_{\max}^{(k)}}{\lambda_{\min}^{(k)}} \leq N(\log(\|\boldsymbol{R}_k\|))^{\frac{1}{3}}, \quad \text{a.s.} \tag{5}$$

where $\lambda_{\max}^{(k)}$ and $\lambda_{\min}^{(k)}$ represent the maximum and minimum eigenvalues of the matrix, respectively, $(n/m)\boldsymbol{I}_m + \sum_{i=1}^n \sum_{j=1}^k \boldsymbol{\varphi}_j^i \boldsymbol{\varphi}_j^{i\,T}$, and $\|\boldsymbol{R}_k\| \to \infty$ as $k \to \infty$.

Remark 7: We give some illustrations for the necessity and practicality of the above cooperative excitation condition. Consider an extreme case where all the regressor vectors $\boldsymbol{\varphi}_j^i$ are equal to zero. It is clear that Assumption 6 is not satisfied, and the unknown parameter $\boldsymbol{\theta}$ cannot be identified since the observations do not contain any information about the unknown parameter. To estimate $\boldsymbol{\theta}$, we should impose some nonzero information (or excitation) conditions on the regressor vectors $\boldsymbol{\varphi}_j^i$. Therefore, we propose the above cooperative excitation condition to guarantee the convergence of the distributed SG algorithm. For some practical engineering systems, such excitation condition can be satisfied to guarantee the performance of the closed-loop systems. For example, the self-tuning regulator which is widely investigated in the field of adaptive control has many applications in engineering systems such as power system and turbine generator system (see [43], [44]). We can design a dither signal in the self-tuning regulator (cf., [45]) to make Assumption 2.2 be satisfied to deal with the conflict between parameter estimation and control performance.

Remark 8: It can be verified that the i.i.d. signals (by the strong law of large numbers) and the stationary ergodic signals (by the ergodic theorem) have the following property.

(Ergodicity Property): For any $i \in \{1, \ldots, n\}$, the regressor vectors $\boldsymbol{\varphi}_j^i$ satisfy the ergodicity property, i.e., there exists a matrix $\boldsymbol{H}_i$ such that

$$\frac{1}{k}\sum_{j=1}^k \boldsymbol{\varphi}_j^i \boldsymbol{\varphi}_j^{i\,T} \xrightarrow{k \to \infty} \boldsymbol{H}_i, \quad \text{a.s.}$$

Furthermore, if $\sum_{i=1}^n \boldsymbol{H}_i$ is positive definite (cf., [19]), then the ergodicity property implies the PE condition in the multiple sensor case, i.e., $\lambda_{\max}^{(k)}/\lambda_{\min}^{(k)} \leq c_2$ with c_2 being a positive constant, which means that Assumption 6 is satisfied. Hence, Assumption 6 is weaker than the PE condition (see e.g., [32], [33], [34]).

Remark 9: Guo [35] proved that the convergence of the standard SG algorithm (i.e., the weighted matrix in Algorithm 1 is an identity matrix) under the following non-PE condition:

$$\frac{\lambda_{\max}\left(\sum_{j=1}^k \boldsymbol{\varphi}_j^i \boldsymbol{\varphi}_j^{i\,T}\right)}{\lambda_{\min}\left(\sum_{j=1}^k \boldsymbol{\varphi}_j^i \boldsymbol{\varphi}_j^{i\,T}\right)} \leq \tilde{N}\left(\log r_k^i\right)^{\frac{1}{3}}, \quad \text{a.s.} \tag{6}$$

where $r_k^i \to \infty$ as $k \to \infty$. Assumption 6 can be degenerated to the condition (6) when the sensor network is degenerated to a single sensor case. In fact, Assumption 6 can reflect the cooperative effect of multiple sensors in the sense that the estimation task can be still fulfilled by the cooperation of multiple sensors even if any of them cannot (see Example 27 in Section V).

Assumption 10: We assume that the system noise $\{\varepsilon_k^i, i = 1, \ldots, n, k \geq 1\}$ is a martingale difference sequence, that is, $E(\boldsymbol{\Xi}_{k+1}|\mathscr{F}_k) = 0$ with $\mathscr{F}_k = \sigma\{\boldsymbol{\varphi}_j^i, \varepsilon_j^i, i = 1, \ldots, n, j \leq k\}$ and $E(\cdot|\cdot)$ being the conditional mathematical expectation, and there exist constants $c_0 > 0$ and $\varepsilon \in [0, 1)$ (which may depend on ω) such that $E(\|\boldsymbol{\Xi}_{k+1}\|^2|\mathscr{F}_k) \leq c_0\|\boldsymbol{R}_k\|^\varepsilon$ almost surely (a.s.)

It is clear that the i.i.d. zero-mean bounded or Gaussian noise ε_k^i which is independent of the regression signals can satisfy Assumption 10.

III. Convergence of Distributed SG Algorithm

In this section, we will provide the convergence analysis of the proposed distributed SG algorithm.

Let the state transition matrix $\boldsymbol{\Psi}(k, j)$ be recursively defined by

$$\boldsymbol{\Psi}(k+1, j) = (\boldsymbol{I}_{mn} - \mu \boldsymbol{G}_k)\boldsymbol{\Psi}(k, j), \quad \boldsymbol{\Psi}(j, j) = \boldsymbol{I}_{mn}. \tag{7}$$

From the error equation (4), we can see that the analysis of the error $\widetilde{\boldsymbol{\Theta}}_{k+1}$ can be divided into two key steps.

1) Analyzing the properties of the product of random matrices $\boldsymbol{\Psi}(k, j) = \prod_{p=j}^{k-1}(\boldsymbol{I}_{mn} - \mu \boldsymbol{G}_p)$.
2) Analyzing the cumulative effect of noises.

To establish the properties of $\boldsymbol{\Psi}(k, j)$, we first show that for small step sizes μ and ν, we have $0 \leq \mu \boldsymbol{G}_k \leq \boldsymbol{I}_{mn}$, which will be used in the convergence analysis of the proposed algorithm.

Lemma 11: Suppose that Assumption 5 is satisfied. If $\mu(1+4\nu) \leq 1$, then we have

$$0 \leq \mu \boldsymbol{G}_k \leq \boldsymbol{I}_{mn}.$$

Proof: By Step 2 in Algorithm 1, we have

$$x_k^i(Q) = \sum_{j=1}^n a_{ij}^{(Q)} \frac{\left\|\boldsymbol{\varphi}_k^j\right\|^2}{r_k^j} \tag{8}$$

where $a_{ij}^{(Q)}$ is the ith row, jth column element of $\mathcal{A}^Q$ (i.e., $\mathcal{A}$ to the power of Q). The matrix $\mathcal{A}$ is stochastic, so is the matrix $\mathcal{A}^Q$ for $Q \geq 1$. Hence, we have for $i \in \{1, \ldots, n\}$

$$x_k^i(Q) \leq \left(\max_{1\leq j\leq n} \frac{\left\|\boldsymbol{\varphi}_k^j\right\|^2}{r_k^j}\right)\sum_{j=1}^n a_{ij}^{(Q)} = \max_{1\leq j\leq n} \frac{\left\|\boldsymbol{\varphi}_k^j\right\|^2}{r_k^j} \leq \|\boldsymbol{A}_k\|.$$

By the definition $\boldsymbol{X}_k(Q)$ in Table I, we have $\|\boldsymbol{X}_k(Q) \otimes \boldsymbol{I}_m\| \leq \|\boldsymbol{A}_k\|$. Using Lemma 1, it follows that:

$$\|\mu \boldsymbol{G}_k\| \leq \mu(\|\boldsymbol{A}_k\| + 4\nu\|\boldsymbol{A}_k\|) \leq \mu(1 + 4\nu) \leq 1. \tag{9}$$

This completes the proof of the lemma. ■

By the definition of $\boldsymbol{G}_k$ in Table I, we see that $\boldsymbol{G}_k$ is a nonnegative definite matrix. Thus, there exists a matrix sequence $\{\boldsymbol{B}_k, k \geq 0\}$, such that for all k we have

$$\boldsymbol{B}_k^2 = \mu \boldsymbol{G}_k. \tag{10}$$

In the following, we will analyze the properties of $\boldsymbol{\Psi}(k, j)$.

Lemma 12: Assume that the step sizes μ and ν satisfy $\mu(1 + 4\nu) \leq 1$. Then for any $k \geq 0$ the following inequality holds:

$$\sum_{j=0}^{k-1} \|\boldsymbol{\Psi}(k, j+1)\boldsymbol{B}_j\|^2 \leq mn.$$

Proof: By the definition of the state transition matrix in (7), we have $\boldsymbol{\Psi}(k, k) = \boldsymbol{I}_{mn}$ and

$$\boldsymbol{\Psi}(k, j+1)\boldsymbol{\Psi}(j+1, j) = \boldsymbol{\Psi}(k, j).$$

Then

$$\begin{aligned} mn &= \operatorname{tr}\Big(\boldsymbol{\Psi}(k, k)\boldsymbol{\Psi}^T(k, k)\Big) \\ &\geq \operatorname{tr}\left(\sum_{j=0}^{k-1}\Big[\boldsymbol{\Psi}(k, j+1)\boldsymbol{\Psi}^T(k, j+1) \right. \\ &\qquad \left. - \boldsymbol{\Psi}(k, j)\boldsymbol{\Psi}^T(k, j)\Big]\right) \\ &= \operatorname{tr}\left(\sum_{j=0}^{k-1}\boldsymbol{\Psi}(k, j+1)\Big[\boldsymbol{I}_{mn} - \boldsymbol{\Psi}(j+1, j) \right. \\ &\qquad \left. \cdot \boldsymbol{\Psi}^T(j+1, j)\Big]\boldsymbol{\Psi}^T(k, j+1)\right). \end{aligned}$$

Furthermore, by $\boldsymbol{\Psi}(j+1, j) = \boldsymbol{I}_{mn} - \mu \boldsymbol{G}_j$, we have

$$\begin{aligned} mn &\geq \operatorname{tr}\left(\sum_{j=0}^{k-1}\boldsymbol{\Psi}(k, j+1)\Big[\mu \boldsymbol{G}_j + \mu \boldsymbol{G}_j(\boldsymbol{I}_{mn} - \mu \boldsymbol{G}_j)\Big] \right. \\ &\qquad \left. \cdot \boldsymbol{\Psi}^T(k, j+1)\right) \\ &\geq \operatorname{tr}\left(\sum_{j=0}^{k-1}\boldsymbol{\Psi}(k, j+1)\mu \boldsymbol{G}_j\boldsymbol{\Psi}^T(k, j+1)\right) \\ &= \operatorname{tr}\left(\sum_{j=0}^{k-1}\boldsymbol{\Psi}(k, j+1)\boldsymbol{B}_j^2\boldsymbol{\Psi}^T(k, j+1)\right) \\ &\geq \sum_{j=0}^{k-1}\|\boldsymbol{\Psi}(k, j+1)\boldsymbol{B}_j\|^2 \end{aligned} \tag{11}$$

which completes the proof of the lemma. ■

How to deal with the noise effect of the distributed SG algorithm is a crucial step for the convergence analysis of the algorithm. The following lemma provides an upper bound of the cumulative summation of the noises.

Lemma 13: Suppose that Assumption 10 is satisfied, and the condition number of $\boldsymbol{R}_k$ is bounded (i.e., there exists a positive constant γ which may depend on the sample ω such that $\max_{1\leq i\leq n} r_k^i / \min_{1\leq i\leq n} r_k^i \leq \gamma$), then $\boldsymbol{S}_k$ tends to a finite limit $\boldsymbol{S}$ as $k \to \infty$, where $\boldsymbol{S}_k \triangleq \sum_{j=0}^{k} \boldsymbol{\Phi}_j \boldsymbol{R}_j^{-1} \boldsymbol{\Xi}_{j+1}^T$. Furthermore, there exists a positive constant c which may depend on the sample ω such that

$$\|\widetilde{\boldsymbol{S}}_{k-1}\| \leq c\|\boldsymbol{R}_k\|^{-\delta} \tag{12}$$

where $\widetilde{\boldsymbol{S}}_{k-1} \triangleq \boldsymbol{S} - \boldsymbol{S}_{k-1}$ and $\delta \in (0, ((1-\varepsilon)/2))$ with ε defined in Assumption 10.

The proof is put in Appendix A.

Now, we present a necessary and sufficient condition for the strong consistency of the distributed SG algorithm.

Theorem 14: Suppose that the condition number of $\boldsymbol{R}_k$ is bounded, and $\mu(1 + 4\nu) < 1$. Then under Assumptions 5 and 10, the estimate $\hat{\boldsymbol{\Theta}}_k$ defined in Table I converges to the true parameter $\boldsymbol{\Theta}$ a.s. for any initial value $\hat{\boldsymbol{\Theta}}_0$ if and only if $\boldsymbol{\Psi}(k, 0) \to 0$ a.s. as $k \to \infty$.

Proof: By (4) and (7), we have the following expression:

$$\begin{aligned} \widetilde{\boldsymbol{\Theta}}_{k+1} &= \boldsymbol{\Psi}(k+1, 0)\widetilde{\boldsymbol{\Theta}}_0 \\ &\quad - \mu \sum_{j=0}^{k} \boldsymbol{\Psi}(k+1, j+1)\boldsymbol{\Phi}_j \boldsymbol{R}_j^{-1} \boldsymbol{\Xi}_{j+1}^T. \end{aligned} \tag{13}$$

We note that the second term on the right-hand side of (13) is independent of $\widetilde{\boldsymbol{\Theta}}_0$. Thus, $\widetilde{\boldsymbol{\Theta}}_{k+1} \to 0$ for any $\widetilde{\boldsymbol{\Theta}}_0$ implies $\boldsymbol{\Psi}(k+1, 0)\widetilde{\boldsymbol{\Theta}}_0 \to 0$ as $k \to \infty$, which means that $\boldsymbol{\Psi}(k+1, 0) \to 0$ as $k \to \infty$. This completes the proof of the necessity part of the theorem.

Now, let us move on to the sufficiency part. It is clear that to prove the convergence of the algorithm, we just need to prove

$$\sum_{j=0}^{k} \boldsymbol{\Psi}(k+1, j+1)\boldsymbol{\Phi}_j \boldsymbol{R}_j^{-1} \boldsymbol{\Xi}_{j+1}^T \to 0 \quad \text{a.s., as } k \to \infty. \tag{14}$$

Set $\boldsymbol{S}_{-1} = 0$. By the definition of $\boldsymbol{\Psi}(\cdot, \cdot)$ in (7), we have

$$\begin{aligned} &\left\|\sum_{j=0}^{k} \boldsymbol{\Psi}(k+1, j+1)\boldsymbol{\Phi}_j \boldsymbol{R}_j^{-1} \boldsymbol{\Xi}_{j+1}^T\right\| \\ &\quad = \left\|\sum_{j=0}^{k} \boldsymbol{\Psi}(k+1, j+1)(\boldsymbol{S}_j - \boldsymbol{S}_{j-1})\right\| \\ &\quad = \left\|\boldsymbol{S}_k - \sum_{j=0}^{k}[\boldsymbol{\Psi}(k+1, j+1) - \boldsymbol{\Psi}(k+1, j)]\boldsymbol{S}_{j-1}\right\| \\ &\quad = \left\|\boldsymbol{S}_k - \sum_{j=0}^{k}[\boldsymbol{\Psi}(k+1, j+1) - \boldsymbol{\Psi}(k+1, j)]\boldsymbol{S} \right. \\ &\qquad \left. + \sum_{j=0}^{k}[\boldsymbol{\Psi}(k+1, j+1) - \boldsymbol{\Psi}(k+1, j)]\widetilde{\boldsymbol{S}}_{j-1}\right\| \\ &\quad = \left\|\boldsymbol{S}_k - \boldsymbol{S} + \boldsymbol{\Psi}(k+1, 0)\boldsymbol{S} + \sum_{j=0}^{k} \boldsymbol{\Psi}(k+1, j+1)\right. \end{aligned}$$

$$\cdot [\boldsymbol{I}_{mn} - \boldsymbol{\Psi}(j+1, j)]\widetilde{\boldsymbol{S}}_{j-1} \Bigg\| \tag{15}$$

where $\boldsymbol{S}, \boldsymbol{S}_k, \widetilde{\boldsymbol{S}}_{k-1}$ are defined in Lemma 13.

Uisng Lemma 13, we have $\boldsymbol{S}_k - \boldsymbol{S} \to 0$. By the condition that $\boldsymbol{\Psi}(k+1, 0) \to 0$ as $k \to \infty$, we have $\boldsymbol{\Psi}(k+1, 0)\boldsymbol{S} \to 0$ as $k \to \infty$. Using Lemma 13 and Hölder inequality, we have

$$\begin{aligned}
&\left\| \sum_{j=0}^{k} \boldsymbol{\Psi}(k+1, j+1)[\boldsymbol{I}_{mn} - \boldsymbol{\Psi}(j+1, j)]\widetilde{\boldsymbol{S}}_{j-1} \right\| \\
&= \left\| \sum_{j=0}^{k} \boldsymbol{\Psi}(k+1, j+1)\mu \boldsymbol{G}_j \widetilde{\boldsymbol{S}}_{j-1} \right\| \\
&\le c \sum_{j=0}^{M} \|\boldsymbol{\Psi}(k+1, j+1)\boldsymbol{B}_j\| \frac{\|\boldsymbol{B}_j\|}{\|\boldsymbol{R}_j\|^{\delta}} \\
&\quad + c \left(\sum_{j=M+1}^{k} \|\boldsymbol{\Psi}(k+1, j+1)\boldsymbol{B}_j\|^2 \right)^{\frac{1}{2}} \\
&\quad \cdot \left(\sum_{j=M+1}^{k} \frac{\|\boldsymbol{B}_j\|^2}{\|\boldsymbol{R}_j\|^{2\delta}} \right)^{\frac{1}{2}}.
\end{aligned} \tag{16}$$

Furthermore, using Lemma 2 and Lemma 11, we have

$$\begin{aligned}
\sum_{j=1}^{\infty} \frac{\|\boldsymbol{B}_j\|^2}{\|\boldsymbol{R}_j\|^{2\delta}} &= \sum_{j=1}^{\infty} \frac{\|\mu \boldsymbol{G}_j\|}{\|\boldsymbol{R}_j\|^{2\delta}} \\
&\le \mu(1+4\nu) \sum_{j=1}^{\infty} \frac{\|\boldsymbol{A}_j\|}{\|\boldsymbol{R}_j\|^{2\delta}} \\
&\le \sum_{j=1}^{\infty} \frac{\|\boldsymbol{\Phi}_j\|^2 \|\boldsymbol{R}_j^{-1}\|}{\|\boldsymbol{R}_j\|^{2\delta}} \le \gamma \sum_{j=1}^{\infty} \frac{\|\boldsymbol{\Phi}_j\|^2}{\|\boldsymbol{R}_j\|^{1+2\delta}} \\
&< \infty.
\end{aligned} \tag{17}$$

According to (17) and Lemma 12, we see that the two terms on the right-hand side of (16) tend to zero if we first let $k \to \infty$, and then let $M \to \infty$. Hence, (14) holds. This completes the proof of the theorem. ■

A key problem still remains unresolved: what conditions on the regression signals $\{\boldsymbol{\varphi}_k^i\}$ can guarantee that $\boldsymbol{\Psi}(k, 0) \to 0$ as $k \to \infty$? In the following, we will prove that under the cooperative excitation condition (i.e., Assumption 6) the convergence results for the distributed algorithm can be established.

Before stating the main theorem of this section, we first give a lemma which provides a key step for the convergence of the matrix $\boldsymbol{\Psi}(k, 0)$.

Lemma 15: Suppose that Assumption 5 is satisfied. Then there exists a positive constant σ, such that the following inequality holds for all $t \ge 0$ and all $u \ge 0$:

$$\lambda_{\min}\left(\sum_{k=\hat{g}(t)}^{\hat{g}(t+u)-1} \boldsymbol{G}_k \right) \ge \sigma \lambda_{\min}\left(\sum_{k=\hat{g}(t)}^{\hat{g}(t+u)-1} \sum_{i=1}^{n} \boldsymbol{A}_k^i \right) \tag{18}$$

where $\boldsymbol{A}_k^i \triangleq ((\boldsymbol{\varphi}_k^i)(\boldsymbol{\varphi}_k^i)^T)/r_k^i)$, $\hat{g}(t) \triangleq \max\{k : d_k \le t\}$, and $d_k \triangleq \sum_{i=1}^{n} \sum_{j=K_0}^{k-1} ((\|\boldsymbol{\varphi}_j^i\|^2)/(\mathrm{tr}(\boldsymbol{R}_j)(\log \mathrm{tr}(\boldsymbol{R}_{j-1}))^{(1/3)}))$, and K_0 is defined in Assumption 6.

Proof: Set

$$\boldsymbol{H}_t^i = \sum_{k=\hat{g}(t)}^{\hat{g}(t+u)-1} \boldsymbol{A}_k^i, \quad \boldsymbol{H}_t = \mathrm{diag}\{\boldsymbol{H}_t^1, \ldots, \boldsymbol{H}_t^n\}$$

$$\boldsymbol{\Delta}_t = \sum_{k=\hat{g}(t)}^{\hat{g}(t+u)-1} \boldsymbol{G}_k, \quad \boldsymbol{\Gamma}_t = \sum_{i=1}^{n} \boldsymbol{H}_t^i.$$

By the definition of $\boldsymbol{A}_k, \boldsymbol{G}_k$ in Table I, we have

$$\boldsymbol{\Delta}_t = \boldsymbol{H}_t + \nu \sum_{k=\hat{g}(t)}^{\hat{g}(t+u)-1} \mathscr{L}(\boldsymbol{X}_k(Q) \otimes \boldsymbol{I}_m)\mathscr{L}$$

$$\boldsymbol{\Gamma}_t = \sum_{k=\hat{g}(t)}^{\hat{g}(t+u)-1} \sum_{i=1}^{n} \boldsymbol{A}_k^i.$$

The eigenvalues of $\mathscr{L}$ are denoted in a nondecreasing order as $l_1, \ldots, l_m, l_{m+1}, \ldots, l_{mn}$, and the corresponding unit orthogonal eigenvectors are denoted as $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_m, \boldsymbol{\xi}_{m+1}, \ldots, \boldsymbol{\xi}_{mn}$. By Assumption 5, we see that $l_1 = \cdots = l_m = 0$ and

$$\boldsymbol{\xi}_1 = \frac{1}{\sqrt{n}}\mathbf{1}_n \otimes \boldsymbol{e}_1, \ldots, \boldsymbol{\xi}_m = \frac{1}{\sqrt{n}}\mathbf{1}_n \otimes \boldsymbol{e}_m \tag{19}$$

where $\mathbf{1}_n$ denotes the n-dimensional vector with all the entries equal to 1, and $\boldsymbol{e}_j$ $(j = 1, \ldots, m)$ is the jth column of the identity matrix $\boldsymbol{I}_m$.

Hence, for any unit vector $\boldsymbol{\eta} \in \mathbb{R}^{mn}$, we have the following expression:

$$\boldsymbol{\eta} = \sum_{j=1}^{m} \kappa_j \boldsymbol{\xi}_j + \sum_{j=m+1}^{mn} \kappa_j \boldsymbol{\xi}_j \triangleq \boldsymbol{\eta}_1 + \boldsymbol{\eta}_2$$

where $\sum_{j=1}^{m} \kappa_j^2 + \sum_{j=m+1}^{mn} \kappa_j^2 = 1$. Then we have

$$\begin{aligned}
\boldsymbol{\eta}^T \boldsymbol{\Delta}_t \boldsymbol{\eta} &= (\boldsymbol{\eta}_1 + \boldsymbol{\eta}_2)^T \left(\boldsymbol{H}_t + \nu \sum_{k=\hat{g}(t)}^{\hat{g}(t+u)-1} \mathscr{L}(\boldsymbol{X}_k(Q) \otimes \boldsymbol{I}_m)\mathscr{L} \right) \\
&\quad \cdot (\boldsymbol{\eta}_1 + \boldsymbol{\eta}_2) \\
&= \boldsymbol{\eta}_1^T \boldsymbol{H}_t \boldsymbol{\eta}_1 + \boldsymbol{\eta}_2^T \boldsymbol{H}_t \boldsymbol{\eta}_2 + 2\boldsymbol{\eta}_1^T \boldsymbol{H}_t \boldsymbol{\eta}_2 \\
&\quad + \nu \boldsymbol{\eta}_1^T \sum_{k=\hat{g}(t)}^{\hat{g}(t+u)-1} \mathscr{L}(\boldsymbol{X}_k(Q) \otimes \boldsymbol{I}_m)\mathscr{L}\boldsymbol{\eta}_1 \\
&\quad + \nu \boldsymbol{\eta}_2^T \sum_{k=\hat{g}(t)}^{\hat{g}(t+u)-1} \mathscr{L}(\boldsymbol{X}_k(Q) \otimes \boldsymbol{I}_m)\mathscr{L}\boldsymbol{\eta}_2 \\
&\quad + 2\nu \boldsymbol{\eta}_1^T \sum_{k=\hat{g}(t)}^{\hat{g}(t+u)-1} \mathscr{L}(\boldsymbol{X}_k(Q) \otimes \boldsymbol{I}_m)\mathscr{L}\boldsymbol{\eta}_2 \\
&\triangleq s_1 + s_2 + s_3 + s_4 + s_5 + s_6.
\end{aligned} \tag{20}$$

In the following, we will estimate $s_i (i = 1, 2, \ldots, 6)$. By the definition of $\boldsymbol{\eta}_1$, we see that $\boldsymbol{\eta}_1$ is the eigenvector corresponding to zero eigenvalue of $\mathscr{L}$. Hence, we have $s_4 = s_6 = 0$.

We note that $\boldsymbol{H}_t$ is a nonnegative definite matrix. Thus, we can decompose it as $\boldsymbol{H}_t = \boldsymbol{H}_t^{(1/2)}\boldsymbol{H}_t^{(1/2)}$, then

$$\begin{aligned} s_3 &= 2\boldsymbol{\eta}_1^T\boldsymbol{H}_t\boldsymbol{\eta}_2 \geq -\zeta\boldsymbol{\eta}_1^T\boldsymbol{H}_t\boldsymbol{\eta}_1 - \frac{1}{\zeta}\boldsymbol{\eta}_2^T\boldsymbol{H}_t\boldsymbol{\eta}_2 \\ &= -\zeta s_1 - \frac{1}{\zeta}s_2 \end{aligned}$$

where the inequality $2\boldsymbol{M}_1^T\boldsymbol{M}_2 \leq \zeta\boldsymbol{M}_1^T\boldsymbol{M}_1 + (1/\zeta)\boldsymbol{M}_2^T\boldsymbol{M}_2$ is used, with ζ being a positive constant and $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ being two matrices with appropriate dimensions.

Let $y = \sum_{j=1}^m \kappa_j^2$. Then by (8) and the definition of $\boldsymbol{X}_k(Q)$, we have

$$\begin{aligned} s_5 &= \nu\boldsymbol{\eta}_2^T \sum_{k=\hat{g}(t)}^{\hat{g}(t+u)-1} \mathscr{L}(\boldsymbol{X}_k(Q)\otimes\boldsymbol{I}_m)\mathscr{L}\boldsymbol{\eta}_2 \\ &\geq \nu \sum_{k=\hat{g}(t)}^{\hat{g}(t+u)-1} \lambda_{\min}(\boldsymbol{X}_k(Q)\otimes\boldsymbol{I}_m) \sum_{j=m+1}^{mn} l_j^2\kappa_j^2 \\ &\geq a\nu \sum_{k=\hat{g}(t)}^{\hat{g}(t+u)-1} \operatorname{tr}(\boldsymbol{A}_k)l_{m+1}^2(1-y) \\ &= a\nu\operatorname{tr}(\boldsymbol{H}_t)l_{m+1}^2(1-y) \end{aligned} \tag{21}$$

where $a \triangleq \min_{i,j\in\{1,\ldots,n\}} a_{ij}^{(Q)}$ is a positive constant for $Q \geq D(\mathscr{G})$ (cf., [46]).

In the following, we will estimate s_1:

$$\begin{aligned} s_1 &= \boldsymbol{\eta}_1^T\boldsymbol{H}_t\boldsymbol{\eta}_1 = \left(\sum_{j=1}^m \kappa_j\boldsymbol{\xi}_j\right)^T \boldsymbol{H}_t \left(\sum_{j=1}^m \kappa_j\boldsymbol{\xi}_j\right) \\ &= \boldsymbol{K}^T\boldsymbol{\Xi}^T\boldsymbol{H}_t\boldsymbol{\Xi}\boldsymbol{K} \end{aligned} \tag{22}$$

where $\boldsymbol{K} = (\kappa_1,\ldots,\kappa_m)^T$ and $\boldsymbol{\Xi} = (\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_m)$. By the definition of $\boldsymbol{\xi}_i$ in (19), we have

$$\boldsymbol{\Xi} = \frac{1}{\sqrt{n}}\begin{pmatrix} \boldsymbol{e}_1 & \boldsymbol{e}_2 & \cdots & \boldsymbol{e}_m \\ \boldsymbol{e}_1 & \boldsymbol{e}_2 & \cdots & \boldsymbol{e}_m \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{e}_1 & \boldsymbol{e}_2 & \cdots & \boldsymbol{e}_m \end{pmatrix}$$

$$\boldsymbol{H}_t\boldsymbol{\Xi} = \frac{1}{\sqrt{n}}\begin{pmatrix} \boldsymbol{H}_t^1\boldsymbol{e}_1 & \boldsymbol{H}_t^1\boldsymbol{e}_2 & \cdots & \boldsymbol{H}_t^1\boldsymbol{e}_m \\ \boldsymbol{H}_t^2\boldsymbol{e}_1 & \boldsymbol{H}_t^2\boldsymbol{e}_2 & \cdots & \boldsymbol{H}_t^2\boldsymbol{e}_m \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{H}_t^n\boldsymbol{e}_1 & \boldsymbol{H}_t^n\boldsymbol{e}_2 & \cdots & \boldsymbol{H}_t^n\boldsymbol{e}_m \end{pmatrix}$$

$$= \frac{1}{\sqrt{n}}\left(\boldsymbol{H}_t^1\ \boldsymbol{H}_t^2,\ldots,\boldsymbol{H}_t^n\right)^T$$

$$\boldsymbol{\Xi}^T\boldsymbol{H}_t\boldsymbol{\Xi} = \frac{1}{n}\begin{pmatrix} \boldsymbol{e}_1^T\boldsymbol{H}_t^1 + \boldsymbol{e}_1^T\boldsymbol{H}_t^2 + \cdots + \boldsymbol{e}_1^T\boldsymbol{H}_t^n \\ \boldsymbol{e}_2^T\boldsymbol{H}_t^1 + \boldsymbol{e}_2^T\boldsymbol{H}_t^2 + \cdots + \boldsymbol{e}_2^T\boldsymbol{H}_t^n \\ \vdots \\ \boldsymbol{e}_m^T\boldsymbol{H}_t^1 + \boldsymbol{e}_m^T\boldsymbol{H}_t^2 + \cdots + \boldsymbol{e}_m^T\boldsymbol{H}_t^n \end{pmatrix}.$$

Hence, we have $\boldsymbol{\Xi}^T\boldsymbol{H}_t\boldsymbol{\Xi} = (1/n)\sum_{i=1}^n \boldsymbol{H}_t^i = (1/n)\boldsymbol{\Gamma}_t$. By this and (22), we have

$$s_1 = \frac{1}{n}\boldsymbol{K}^T\boldsymbol{\Gamma}_t\boldsymbol{K} \geq \frac{\lambda_{\min}(\boldsymbol{\Gamma}_t)}{n}y. \tag{23}$$

By the definition of s_2, we have

$$s_2 = \boldsymbol{\eta}_2^T\boldsymbol{H}_t\boldsymbol{\eta}_2 \leq \operatorname{tr}(\boldsymbol{H}_t)(1-y). \tag{24}$$

Substitute (21)–(24) into (20), we have for $\zeta \in (0,1)$

$$\begin{aligned} \boldsymbol{\eta}^T\boldsymbol{\Delta}_t\boldsymbol{\eta} &\geq (1-\zeta)s_1 + \left(1-\frac{1}{\zeta}\right)s_2 + s_5 \\ &\geq \frac{(1-\zeta)\lambda_{\min}(\boldsymbol{\Gamma}_t)}{n}y + \left(1-\frac{1}{\zeta}\right)\operatorname{tr}(\boldsymbol{H}_t)(1-y) \\ &\quad + a\nu\operatorname{tr}(\boldsymbol{H}_t)l_{m+1}^2(1-y). \end{aligned} \tag{25}$$

Thus, we have

$$\begin{aligned} \lambda_{\min}(\boldsymbol{\Delta}_t) \geq \Bigg[\frac{(1-\zeta)\lambda_{\min}(\boldsymbol{\Gamma}_t)}{n} - \Bigg(a\nu\operatorname{tr}(\boldsymbol{H}_t)l_{m+1}^2 \\ + \operatorname{tr}(\boldsymbol{H}_t) - \frac{\operatorname{tr}(\boldsymbol{H}_t)}{\zeta}\Bigg)\Bigg]y \\ + a\nu\operatorname{tr}(\boldsymbol{H}_t)l_{m+1}^2 + \operatorname{tr}(\boldsymbol{H}_t) - \frac{\operatorname{tr}(\boldsymbol{H}_t)}{\zeta}. \end{aligned}$$

Taking $\zeta = (1/(1+0.5l_{m+1}^2a\nu)) \in (0,1)$, then we can obtain the following inequality:

$$\begin{aligned} \lambda_{\min}(\boldsymbol{\Delta}_t) \geq \Big[\sigma\lambda_{\min}(\boldsymbol{\Gamma}_t) - 0.5a\nu l_{m+1}^2\operatorname{tr}(\boldsymbol{H}_t)\Big]y \\ + 0.5a\nu l_{m+1}^2\operatorname{tr}(\boldsymbol{H}_t) \end{aligned}$$

where $\sigma \triangleq ((l_{m+1}^2a\nu)/(2n + l_{m+1}^2a\nu n)) \in (0,1)$. Hence, by $0 \leq \lambda_{\min}(\boldsymbol{\Gamma}_t) \leq \operatorname{tr}(\boldsymbol{H}_t)$ and $y \in (0,1)$, it is easy to obtain

$$\begin{aligned} &\lambda_{\min}(\boldsymbol{\Delta}_t) - \sigma\lambda_{\min}(\boldsymbol{\Gamma}_t) \\ &\quad \geq \Big[\sigma\lambda_{\min}(\boldsymbol{\Gamma}_t) - 0.5a\nu l_{m+1}^2\operatorname{tr}(\boldsymbol{H}_t)\Big](y-1) > 0. \end{aligned}$$

This completes the proof of the lemma. ■

Remark 16: From Theorem 14, we can see that the properties of the product of the matrices $(I_{mn} - \mu\boldsymbol{G}_j)$, $j \geq 0$ are crucial for the strong consistency of the distributed SG algorithm. Lemma 15 establishes a connection between the eigenvalues of $\{\boldsymbol{A}_k^i\}$ and $\{\boldsymbol{G}_k\}$, and thus builds a bridge between the standard SG algorithm and the distributed SG algorithm.

using Lemma 15, we have the following theorem.

Theorem 17: Let $\mu(1+4\nu) < 1$. Suppose that there exists $i_1 \in \{1,\ldots,n\}$ such that $\limsup_{k\to\infty}(r_k^{i_1}/r_{k-1}^{i_1}) \triangleq r^* < \infty$, and the condition number of $\boldsymbol{R}_k$ is bounded. Under Assumptions 5 and 6, we have $\boldsymbol{\Psi}(k,0) \to 0$ as $k \to \infty$ a.s..

The proof of Theorem 17 is complicated, and we put it in Appendix B.

Remark 18: For some typical cases such as the bounded sequence $\{\boldsymbol{\varphi}_k^i\}$ (i.e., $c' \leq \|\boldsymbol{\varphi}_k^i\| \leq c''$) and the i.i.d. sequence $\{\boldsymbol{\varphi}_k^i\}$, the condition $\limsup_{k\to\infty}(r_k^{i_1}/r_{k-1}^{i_1}) \triangleq r^* < \infty$ of Theorem 17 can be easily verified.

Using Theorems 14 and 17, we have the following corollary.

Corollary 19: Under the conditions of Theorem 17, if Assumption 10 is further satisfied, the convergence of the distributed SG algorithm designed in Algorithm 1 can be obtained.

Proof: Using Theorem 17, we have $\boldsymbol{\Psi}(k,0) \to 0$, as $k \to \infty$. Hence, from Theorem 14, the estimate $\hat{\boldsymbol{\Theta}}_k$ defined in Table I converges to the true parameter $\boldsymbol{\Theta}$ a.s., which completes the proof of the corollary. ■

Although the convergence of the distributed SG algorithm is established for the time-invariant connected graphs, the result is also true when the graphs are time-varying and connected at

each time instant by following the proof line of Corollary 19. For a more general case where graphs are jointly connected, the challenges for the theoretical analysis lie in the influence of the Laplacian matrix of the corresponding graphs on the product of the state transition matrices, and the theoretical investigation for such a case will fall into our future work.

Remark 20: The cooperative excitation condition proposed in Assumption 6 is reasonable in a sense that if there exists only one sensor which satisfies the excitation condition (6), then all the sensors in the network can satisfy Assumption 6. We show this point under a mild condition (the condition number of $\boldsymbol{R}_k$ is bounded) as follows:

$$\frac{\lambda_{\max}^{(k)}}{\lambda_{\min}^{(k)}} \le mn\gamma \cdot \frac{\lambda_{\max}\left\{\sum_{j=1}^{k} \boldsymbol{\varphi}_j^{i_2}\boldsymbol{\varphi}_j^{i_2\,T}\right\}}{\lambda_{\min}\left\{\sum_{j=1}^{k} \boldsymbol{\varphi}_j^{i_2}\boldsymbol{\varphi}_j^{i_2\,T}\right\}} \le mn\gamma\,\tilde{N}\left(\log r_k^{i_2}\right)^{\frac{1}{3}} \le mn\gamma\,\tilde{N}(\log(\|\boldsymbol{R}_k\|))^{\frac{1}{3}}$$

where i_2 denotes the index of the sensor satisfying (6), γ is defined in Lemma 13, and $\lambda_{\max}^{(k)}$ and $\lambda_{\min}^{(k)}$ are defined in Assumption 6.

Combining Remark 20 with Corollary 19, we see that if only one sensor can fulfill the estimation task, then all the sensors in our proposed algorithm can fulfill it.

Remark 21: Different from most results in the literature, our results are obtained without using the independency and stationarity assumptions on the regression signals, which makes it possible to apply the distributed algorithm to practical feedback systems.

IV. Convergence Rate of the Distributed SG Algorithm

In this section, we will consider the convergence rate of the distributed SG algorithm. To prove the theorems of this section, we first introduce the following two lemmas.

Lemma 22: If $\mu(1+4\nu) < 1$, then there exists a constant $\tau_1 \ge 1$ such that for any $k \ge 0$, we have

$$\det(\boldsymbol{I}_{mn} - \mu\boldsymbol{G}_k) \ge [\det(\boldsymbol{I}_{mn} - \boldsymbol{A}_k)]^{\tau_1}.$$

Proof: By the definition of $\boldsymbol{G}_k$ and (9), we have

$$\begin{aligned}\det(\boldsymbol{I}_{mn} - \mu\boldsymbol{G}_k) &\ge (\lambda_{\min}(\boldsymbol{I}_{mn} - \mu\boldsymbol{G}_k))^{mn}\\ &= (1 - \lambda_{\max}(\mu\boldsymbol{G}_k))^{mn}\\ &\ge [1 - \lambda_{\max}(\mu(1+4\nu)\boldsymbol{A}_k)]^{mn}\\ &= [\lambda_{\min}(\boldsymbol{I} - \mu(1+4\nu)\boldsymbol{A}_k)]^{mn}\\ &\ge [\det(\boldsymbol{I} - \mu(1+4\nu)\boldsymbol{A}_k)]^{mn}\\ &\ge [\det(\boldsymbol{I} - \boldsymbol{A}_k)]^{mn}.\end{aligned}$$

The lemma can be proved by taking $\tau_1 = mn$. ■

Lemma 23: If $\mu(1+4\nu) < 1$, then we have the following inequalities,[1]

1) $\|\boldsymbol{\Psi}(k,j)\| \le 1, \quad 0 \le j \le k,\ k \ge 0$

2) $\dfrac{1}{\|\boldsymbol{R}_k\|^{\tau_1}} = O(\|\boldsymbol{\Psi}(k+1,0)\|^m), \quad k \ge 1$

3) $\|\boldsymbol{\Psi}(k,j+1)\| = O(\|\boldsymbol{\Psi}(k,0)\|\|\boldsymbol{R}_j\|^{n\tau_1})$

4) $\displaystyle\sum_{j=M+1}^{\infty} \frac{\|\boldsymbol{\Phi}_j\|^2}{\|\boldsymbol{R}_j\|^{1+\varsigma}} \le \frac{n^{1+\varsigma}}{\varsigma}\frac{1}{\|\boldsymbol{R}_M\|^{\varsigma}}, \quad \varsigma > 0$

where τ_1 is defined in Lemma 22.

The proof of the above lemma is given in Appendix C.

In the following, we establish the specific relationship between the error $\boldsymbol{\Theta}_k$ and $\boldsymbol{\Psi}(k,0)$.

Lemma 24: Under the conditions of Theorem 14, if $\lim_{k\to\infty}\boldsymbol{\Psi}(k,0) = 0$, then we have the following inequality:

$$\|\hat{\boldsymbol{\Theta}}_k - \boldsymbol{\Theta}\| = O\left(\|\boldsymbol{\Psi}(k,0)\|^{\frac{\delta}{n\tau_1(1+\delta)}}\right) \quad \text{a.s.}$$

where δ is defined in Lemma 13 and τ_1 can be taken as mn.

Proof: Let

$$\beta(k) = \max\{j : \|\boldsymbol{R}_j\|^{n\tau_1} \le k\}, \quad k \ge 0$$
$$\bar{\Delta}(k) = \beta\left(\|\boldsymbol{\Psi}(k,0)\|^{-\frac{1}{1+\delta}}\right), \quad k \ge 0.$$

By the definition of $\beta(k)$ and $\bar{\Delta}(k)$, we have

$$\begin{aligned}\|\boldsymbol{R}_{\bar{\Delta}(k)}\|^{n\tau_1} &\le \|\boldsymbol{\Psi}(k,0)\|^{-\frac{1}{1+\delta}}\\ \|\boldsymbol{R}_{\bar{\Delta}(k)+1}\|^{n\tau_1} &> \|\boldsymbol{\Psi}(k,0)\|^{-\frac{1}{1+\delta}}.\end{aligned} \tag{26}$$

According to Lemma 23 3), we have

$$\begin{aligned}\|\boldsymbol{\Psi}(k,\bar{\Delta}(k)+1)\| &= O(\|\boldsymbol{\Psi}(k,0)\|\cdot\|\boldsymbol{R}_{\bar{\Delta}(k)}\|^{n\tau_1})\\ &= O\left(\|\boldsymbol{\Psi}(k,0)\|^{\frac{\delta}{1+\delta}}\right).\end{aligned} \tag{27}$$

We prove the following inequality by contradiction:

$$\bar{\Delta}(k) < k-1, \quad \text{for large } k. \tag{28}$$

Suppose that there exists a large constant k_0 such that $\bar{\Delta}(k_0) \ge k_0 - 1$. Then by (26), we have

$$\|\boldsymbol{R}_{k_0-1}\|^{n\tau_1} \le \|\boldsymbol{\Psi}(k_0,0)\|^{-\frac{1}{1+\delta}}.$$

Using Lemma 23 2), we see that there exists a positive constant $\hat{c}$ such that

$$\|\boldsymbol{R}_{k_0-1}\|^{n\tau_1} \ge \hat{c}\|\boldsymbol{\Psi}(k_0,0)\|^{-mn}.$$

Thus, we have

$$\hat{c} \le \|\boldsymbol{\Psi}(k_0,0)\|^{mn-\frac{1}{1+\delta}}$$

which is contradictory with $\boldsymbol{\Psi}(k,0) \to 0$ as $k \to \infty$.

Note that by (9), we have $\|\mu\boldsymbol{G}_j\| \le \mu(1+4\nu)\gamma((\|\boldsymbol{\Phi}_j\|^2)/(\|\boldsymbol{R}_j\|))$. Hence, from Lemma 13 and (15) in the proof of Theorem 14, we have the following estimation for the noise term of the system:

$$\begin{aligned}&\left\|\sum_{j=0}^{k-1}\boldsymbol{\Psi}(k,j+1)\boldsymbol{\Phi}_j\boldsymbol{R}_j^{-1}\boldsymbol{\Xi}_{j+1}^{T}\right\|\\ &\quad\le \|\widetilde{\boldsymbol{S}}_{k-1}\| + \|\boldsymbol{\Psi}(k,0)\boldsymbol{S}\|\\ &\qquad+\sum_{j=0}^{k-1}\|\boldsymbol{\Psi}(k,j+1)\|\cdot\|\mu\boldsymbol{G}_j\|\cdot\|\widetilde{\boldsymbol{S}}_{j-1}\|\end{aligned}$$

[1]Let $\{\boldsymbol{A}_k\}$ be a matrix sequence and $\{b_k\}$ be a positive scalar sequence. Then by $\boldsymbol{A}_k = O(b_k)$ we mean that there exists a constant $M > 0$ such that $\|\boldsymbol{A}_k\| \le Mb_k,\ \forall k \ge 0$.

$$= O(\|\boldsymbol{R}_k\|^{-\delta}) + O(\|\boldsymbol{\Psi}(k,0)\|) + c\mu(1+4\nu)\gamma \sum_{j=0}^{k-1} \|\boldsymbol{\Psi}(k,j+1)\| \cdot \frac{\|\boldsymbol{\Phi}_j\|^2}{\|\boldsymbol{R}_j\|^{1+\delta}}.$$

Combining this with (28), we have

$$\left\| \sum_{j=0}^{k-1} \boldsymbol{\Psi}(k,j+1)\boldsymbol{\Phi}_j \boldsymbol{R}_j^{-1} \boldsymbol{\Xi}_{j+1}^T \right\| = O(\|\boldsymbol{R}_{\bar{\Delta}(k)+1}\|^{-\delta}) + O(\|\boldsymbol{\Psi}(k,0)\|) + O\left(\sum_{j=0}^{k-1} \|\boldsymbol{\Psi}(k,j+1)\| \cdot \frac{\|\boldsymbol{\Phi}_j\|^2}{\|\boldsymbol{R}_j\|^{1+\delta}} \right). \tag{29}$$

In the following, we estimate the last term on the right-hand side of (29). Using Lemma 23 1) and (27), we have

$$\begin{aligned} &\sum_{j=0}^{k-1} \|\boldsymbol{\Psi}(k,j+1)\| \cdot \frac{\|\boldsymbol{\Phi}_j\|^2}{\|\boldsymbol{R}_j\|^{1+\delta}} \\ &\le \sum_{j=0}^{\bar{\Delta}(k)} \|\boldsymbol{\Psi}(k,\bar{\Delta}(k)+1)\| \cdot \|\boldsymbol{\Psi}(\bar{\Delta}(k)+1,j+1)\| \cdot \frac{\|\boldsymbol{\Phi}_j\|^2}{\|\boldsymbol{R}_j\|^{1+\delta}} + \sum_{j=\bar{\Delta}(k)+1}^{k-1} \|\boldsymbol{\Psi}(k,j+1)\| \cdot \frac{\|\boldsymbol{\Phi}_j\|^2}{\|\boldsymbol{R}_j\|^{1+\delta}} \\ &= O\left(\|\boldsymbol{\Psi}(k,0)\|^{\frac{\delta}{1+\delta}}\right) \sum_{j=0}^{\infty} \frac{\|\boldsymbol{\Phi}_j\|^2}{\|\boldsymbol{R}_j\|^{1+\delta}} + \sum_{j=\bar{\Delta}(k)+1}^{k-1} \frac{\|\boldsymbol{\Phi}_j\|^2}{\|\boldsymbol{R}_j\|^{1+\delta}}. \end{aligned} \tag{30}$$

Furthermore, using Lemma 2 and Lemma 23 4), we have

$$\begin{aligned} &\sum_{j=0}^{k-1} \|\boldsymbol{\Psi}(k,j+1)\| \cdot \frac{\|\boldsymbol{\Phi}_j\|^2}{\|\boldsymbol{R}_j\|^{1+\delta}} \\ &= O\left(\|\boldsymbol{\Psi}(k,0)\|^{\frac{\delta}{1+\delta}}\right) + \frac{\|\boldsymbol{\Phi}_{\bar{\Delta}(k)+1}\|^2}{\|\boldsymbol{R}_{\bar{\Delta}(k)+1}\|^{1+\delta}} + \sum_{j=\bar{\Delta}(k)+2}^{\infty} \frac{\|\boldsymbol{\Phi}_j\|^2}{\|\boldsymbol{R}_j\|^{1+\delta}} \\ &= O\left(\|\boldsymbol{\Psi}(k,0)\|^{\frac{\delta}{1+\delta}}\right) + O(\|\boldsymbol{R}_{\bar{\Delta}(k)+1}\|^{-\delta}). \end{aligned}$$

Combining the above inequality with (26) and (29), then we have

$$\begin{aligned} &\left\| \sum_{j=0}^{k-1} \boldsymbol{\Psi}(k,j+1)\boldsymbol{\Phi}_j \boldsymbol{R}_j^{-1} \boldsymbol{\Xi}_{j+1}^T \right\| \\ &= O(\|\boldsymbol{R}_{\bar{\Delta}(k)+1}\|^{-\delta}) + O(\|\boldsymbol{\Psi}(k,0)\|) + O\left(\|\boldsymbol{\Psi}(k,0)\|^{\frac{\delta}{1+\delta}}\right) \\ &= O\left(\|\boldsymbol{\Psi}(k,0)\|^{\frac{\delta}{n\tau_1(1+\delta)}}\right) + O(\|\boldsymbol{\Psi}(k,0)\|) + O\left(\|\boldsymbol{\Psi}(k,0)\|^{\frac{\delta}{1+\delta}}\right) \\ &= O\left(\|\boldsymbol{\Psi}(k,0)\|^{\frac{\delta}{n\tau_1(1+\delta)}}\right) \end{aligned} \tag{31}$$

where $\|\boldsymbol{\Psi}(k,0)\| \le 1$ is used in the above inequality. Hence by (13) and (31), we have

$$\|\widetilde{\boldsymbol{\Theta}}_k\| = O\left(\|\boldsymbol{\Psi}(k,0)\|^{\frac{\delta}{n\tau_1(1+\delta)}}\right) \quad \text{a.s.} \tag{32}$$

This completes the proof of the lemma. ■

We further establish the convergence rate under the cooperative excitation condition (Assumption 6). We see that the rate can be expressed via simply characterizable quantities.

Theorem 25: Under the conditions of Theorem 17 and Assumption 10, we have

$$\|\widetilde{\boldsymbol{\Theta}}_k\| = O\left((\log\|\boldsymbol{R}_k\|)^{-\delta_1}\right) \quad \text{a.s.}$$

with δ_1 being a positive constant.

The proof of the above theorem is given in Appendix D.

Remark 26: Using Theorem 25, the convergence rate of the proposed distributed SG algorithm is mainly determined by the parameters δ_1 and $\|\boldsymbol{R}_k\|$. After a careful calculation by following the proof of Theorem 25 in Appendix D, we can obtain that the parameter δ_1 is positively related to μ, ν, and l_{m+1}, where l_{m+1} is the $(m+1)$th eigenvalue of the Laplacian matrix $\mathscr{L} = \boldsymbol{L} \otimes \boldsymbol{I}_m$ in a nondecreasing order and can be regarded as a measure of the connectivity of the graph $\mathscr{G}$.

V. Simulation Example

In this section, we provide an example of nonindependent regression signals to illustrate the cooperative effect in this article.

Example 27: The network is composed of $n = 28$ sensors whose dynamics obey (1) with $m = 10$. The system noises ε_k^i, $(i = 1,\ldots,n,\ k \ge 1)$ in (1) are independent and identically distributed with $\varepsilon_k^i \sim \mathcal{N}(0, 1.2^2)$ (Gaussian distribution with zero mean and variance 1.2^2). Set the unknown parameter as $\boldsymbol{\theta} = [1,2,\ldots,10]^T$. Let $\boldsymbol{\varphi}_k^i \in \mathbb{R}^{10}$ $(i = 1,\ldots,n,\ k \ge 1)$ be generated by a state-space model

$$\begin{aligned} \boldsymbol{u}_k^i &= \boldsymbol{A}_i \boldsymbol{u}_{k-1}^i + \boldsymbol{B}_i \xi_k^i \\ \boldsymbol{\varphi}_k^i &= \boldsymbol{C}_i \boldsymbol{u}_k^i. \end{aligned} \tag{33}$$

For convenience, we take

$$\begin{aligned} \boldsymbol{A}_i &= \operatorname{diag}\{1.2,\ldots,1.2\} \in \mathbb{R}^{10\times 10} \\ \boldsymbol{B}_i &= \boldsymbol{e}_j \in \mathbb{R}^{10} \\ \boldsymbol{C}_i &= \operatorname{col}\{0,\ldots,0,\underset{j^{\text{th}}}{\boldsymbol{e}_j^T},0,\ldots,0\} \in \mathbb{R}^{10\times 10} \end{aligned}$$

where $j = \mod(i,m)$ and $\boldsymbol{e}_j$ $(j = 1,\ldots,m)$ is the jth column of the identity matrix $\boldsymbol{I}_m$ $(m = 10)$. Let $\{\xi_k^i, k \ge 1, i = 1,\ldots,n\}$ be independent and identically distributed with $\xi_k^i \in \mathbb{R} \sim \mathcal{N}(0, 0.3^2)$.

The network structure is shown in Fig. 1. Here, we use the Metropolis rule in [47] to construct the weights.

We can verify that for each sensor $i (i = 1,\ldots,28)$, the regression signals $\boldsymbol{\varphi}_k^i$ generated by (33) cannot satisfy the excitation condition (6), but they can cooperatively satisfy Assumption 6. We repeat the simulation for $s = 500$ times with the same initial states.

1) For fixed step sizes $\mu = 0.25$ and $\nu = 0.7$, the simulation results are shown in Fig. 2. We see

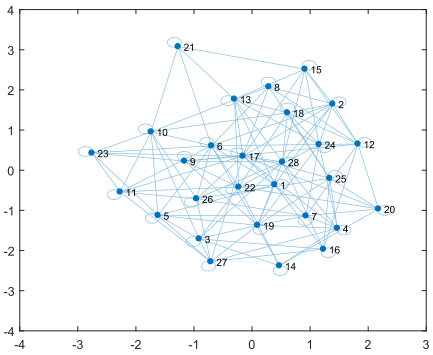


Fig. 1. Network topology.

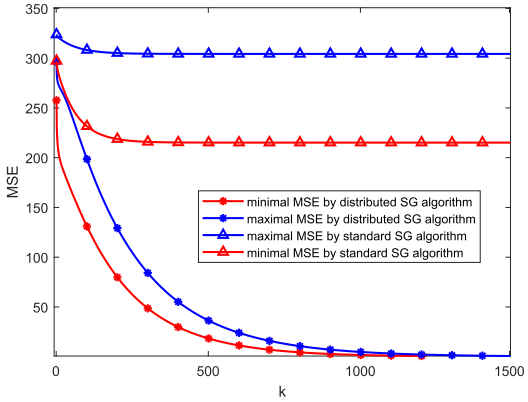


Fig. 2. Maximal and minimal MSEs of sensors using the standard SG algorithm and the distributed SG algorithm.

that if all the sensors use the standard SG algorithm (i.e., the weight matrix is an identity matrix) to estimate $\boldsymbol{\theta}$, the mean square error (MSE) of each sensor [i.e.,$(1/s)\sum_{p=1}^{s}\|\boldsymbol{\theta}_k^{i,p}-\boldsymbol{\theta}\|^2, k=1,2,\ldots,$ where the superscript p being the pth simulation.] cannot converge to zero, while the MSE of each sensor in the distributed SG algorithm (i.e., Algorithm 1) converges to zero. It is clear that the estimation task can be fulfilled through exchanging information between sensors even though any individual sensor cannot.

2) We compare our algorithm (Algorithm 1) with the distributed algorithms in [23] and [24] which are obtained by the distributed optimization method.
(Distributed SG in [23])

$$\boldsymbol{\theta}_{k+1}^{i}=\boldsymbol{\theta}_k^i-\alpha_k\sum_{l\in N^i}\left(\boldsymbol{\theta}_k^i-\boldsymbol{\theta}_k^l\right)+\beta_k\frac{\boldsymbol{\varphi}_k^i}{r_k^i}\left[y_{k+1}^i-\left(\boldsymbol{\varphi}_k^i\right)^T\left(\boldsymbol{\theta}_k^i+\boldsymbol{\zeta}_k^i\right)\right]+\gamma_k\boldsymbol{\zeta}_k'^i$$

where $\alpha_k=(1/k^{(1/3)})$, $\beta_k=(1/k)$, $\gamma_k=(1/(k^{(1/2)}(\log\log k)^{1/2}))$, $\boldsymbol{\zeta}_k^i$ is the gradient measurement noise, and $\boldsymbol{\zeta}_k'^i$ is the annealing noise.
(Distributed SG in [24])

$$\boldsymbol{\theta}_{k+1}^{i}=\boldsymbol{\theta}_k^i-\eta_k\sum_{l\in N^i}\left(\boldsymbol{\theta}_k^i-\boldsymbol{\theta}_k^l\right)+\lambda_k\left[\frac{\boldsymbol{\varphi}_k^i}{r_k^i}\left(y_{k+1}^i-\left(\boldsymbol{\varphi}_k^i\right)^T\boldsymbol{\theta}_k^i\right)-\boldsymbol{\zeta}_k''^i\right]$$

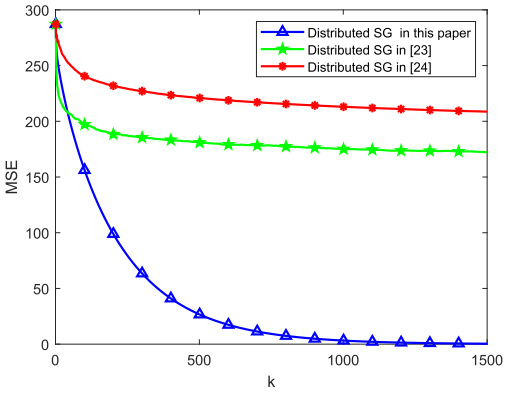


Fig. 3. Comparison of different distributed SG algorithms.

where $\eta_k=(1/(28(1+k)^{(1/3)}))$ is the weight, $\lambda_k=(1/(1+k))$ is the step size, and $\boldsymbol{\zeta}_k''^i$ is the noise.

We conduct the simulations using the same regressors, initial states, and step sizes as above. The average MSEs on the whole network [i.e.,$(1/ns)\sum_{i=1}^{n}\sum_{p=1}^{s}\|\boldsymbol{\theta}_k^{i,p}-\boldsymbol{\theta}\|^2, k=1,2,\ldots$] of the different distributed algorithms are shown in Fig. 3, from which we see that the algorithm proposed in this article has a faster convergence rate than the algorithms in [23] and [24].

VI. Conclusion

This article proposed a distributed SG algorithm based on the consensus strategy and the diffusion of the regression vectors to cooperatively estimate an unknown time-invariant parameter. We introduced a cooperative excitation condition, under which the almost sure convergence of the proposed algorithm can be guaranteed, and the convergence rate of the algorithm can be established. Compared with the existing results concerning the distributed estimation in the literature, our results are obtained without relying on the independency and stationarity assumptions, which makes it possible to apply our results to the feedback control systems. Furthermore, we found that the sensors can cooperate to finish the estimation task even though any individual cannot. Many interesting problems deserve to be further investigated, for example, the convergence of the distributed SG algorithm with correlated noise, the analysis of other distributed algorithms such as the distributed Kalman filter, and the combination of the distributed adaptive estimation with the distributed control.

Appendix A
Proof of Lemma 13

Proof: By $\varepsilon<1$, we see that $2-\varepsilon-2\delta>1$. Since $\boldsymbol{R}_k=\boldsymbol{R}_{k-1}+\boldsymbol{\Phi}_k^T\boldsymbol{\Phi}_k$, we obtain

$$\operatorname{tr}(\boldsymbol{R}_k)=\operatorname{tr}(\boldsymbol{R}_{k-1})+\operatorname{tr}\left(\boldsymbol{\Phi}_k^T\boldsymbol{\Phi}_k\right). \tag{34}$$

Then using Lemma 2, we have the following inequality:

$$\sum_{k=1}^{\infty}\frac{\|\boldsymbol{\Phi}_k\|^2}{\|\boldsymbol{R}_k\|^{2-2\delta-\varepsilon}}\le n^{2-2\delta-\varepsilon}\sum_{k=1}^{\infty}\frac{\operatorname{tr}\left(\boldsymbol{\Phi}_k^T\boldsymbol{\Phi}_k\right)}{(\operatorname{tr}(\boldsymbol{R}_k))^{2-2\delta-\varepsilon}}<\infty.$$

By the boundedness of the condition number of $\boldsymbol{R}_k$, we have $\|\boldsymbol{R}_k\|\|\boldsymbol{R}_k^{-1}\| \le \gamma$ for all k. Using this and the above inequality, we have

$$\begin{aligned}
&\sum_{k=1}^{\infty} E\left(\left\|\boldsymbol{\Phi}_k \boldsymbol{R}_k^{\delta-1}\boldsymbol{\Xi}_{k+1}^T\right\|^2 \middle| \mathscr{F}_k\right) \\
&\quad \le c_0 \sum_{k=1}^{\infty} \|\boldsymbol{\Phi}_k\|^2 \left\|\boldsymbol{R}_k^{\delta-1}\right\|^2 \|\boldsymbol{R}_k\|^{\varepsilon} \\
&\quad = c_0 \sum_{k=1}^{\infty} \|\boldsymbol{\Phi}_k\|^2 \left\|\boldsymbol{R}_k^{-1}\right\|^{2-2\delta} \|\boldsymbol{R}_k\|^{\varepsilon} \\
&\quad \le c_0 \gamma^{2-2\delta} \sum_{k=1}^{\infty} \frac{\|\boldsymbol{\Phi}_k\|^2}{\|\boldsymbol{R}_k\|^{2-2\delta-\varepsilon}} < \infty, \quad \text{a.s.}
\end{aligned}$$

where Assumption 10 is used in the first inequality. By the martingale convergence theorem, it follows that as $k \to \infty$, $\sum_{j=1}^{k} \boldsymbol{\Phi}_j \boldsymbol{R}_j^{\delta-1}\boldsymbol{\Xi}_{j+1}^T$ converges a.s. Hence for any $\eta > 0$, if k is large enough, then we have $\|\widetilde{\boldsymbol{S}}_{k-1,\delta}\| < \eta$, where $\widetilde{\boldsymbol{S}}_{k-1,\delta} \triangleq \sum_{j=k}^{\infty} \boldsymbol{\Phi}_j \boldsymbol{R}_j^{\delta-1}\boldsymbol{\Xi}_{j+1}^T$.

By the definition of $\boldsymbol{\Phi}_j$ and $\boldsymbol{R}_j$ in Table I, we see that $\boldsymbol{\Phi}_j \boldsymbol{R}_j^{-\delta} = (\boldsymbol{R}_j^{-\delta} \otimes \boldsymbol{I}_m)\boldsymbol{\Phi}_j$. Then summation by parts yields the following result:

$$\begin{aligned}
&\left\|\left(\boldsymbol{R}_k^{\delta} \otimes \boldsymbol{I}_m\right)\widetilde{\boldsymbol{S}}_{k-1}\right\| \\
&\quad = \left\|\left(\boldsymbol{R}_k^{\delta} \otimes \boldsymbol{I}_m\right)\sum_{j=k}^{\infty}\left(\boldsymbol{R}_j^{-\delta} \otimes \boldsymbol{I}_m\right)\boldsymbol{\Phi}_j \boldsymbol{R}_j^{\delta-1}\boldsymbol{\Xi}_{j+1}^T\right\| \\
&\quad = \left\|\left(\boldsymbol{R}_k^{\delta} \otimes \boldsymbol{I}_m\right)\sum_{j=k}^{\infty}\left(\boldsymbol{R}_j^{-\delta} \otimes \boldsymbol{I}_m\right)\left(\widetilde{\boldsymbol{S}}_{j-1,\delta} - \widetilde{\boldsymbol{S}}_{j,\delta}\right)\right\| \\
&\quad = \left\|\widetilde{\boldsymbol{S}}_{k-1,\delta} - \left(\boldsymbol{R}_k^{\delta} \otimes \boldsymbol{I}_m\right)\sum_{j=k}^{\infty}\left(\left(\boldsymbol{R}_j^{-\delta} - \boldsymbol{R}_{j+1}^{-\delta}\right) \otimes \boldsymbol{I}_m\right)\widetilde{\boldsymbol{S}}_{j,\delta}\right\| \\
&\quad \le \eta + \eta\left\|\left(\boldsymbol{R}_k^{\delta} \otimes \boldsymbol{I}_m\right)\sum_{j=k}^{\infty}\left(\left(\boldsymbol{R}_j^{-\delta} - \boldsymbol{R}_{j+1}^{-\delta}\right) \otimes \boldsymbol{I}_m\right)\right\| \\
&\quad \le 2\eta. \qquad (35)
\end{aligned}$$

Furthermore, by (35) we have

$$\begin{aligned}
\|\widetilde{\boldsymbol{S}}_{k-1}\| &= \left\|\left(\boldsymbol{R}_k^{-\delta} \otimes \boldsymbol{I}_m\right)\left(\boldsymbol{R}_k^{\delta} \otimes \boldsymbol{I}_m\right)\widetilde{\boldsymbol{S}}_{k-1}\right\| \\
&\le \left\|\left(\boldsymbol{R}_k^{-\delta} \otimes \boldsymbol{I}_m\right)\right\|\left\|\left(\boldsymbol{R}_k^{\delta} \otimes \boldsymbol{I}_m\right)\widetilde{\boldsymbol{S}}_{k-1}\right\| \\
&\le 2\eta\left\|\left(\boldsymbol{R}_k^{-1} \otimes \boldsymbol{I}_m\right)\right\|^{\delta} = 2\eta\left\|\boldsymbol{R}_k^{-1}\right\|^{\delta} \le \frac{2\eta\gamma^{\delta}}{\|\boldsymbol{R}_k\|^{\delta}}
\end{aligned}$$

where $\|\boldsymbol{R}_k^{-1} \otimes \boldsymbol{I}_m\| = \|\boldsymbol{R}_k^{-1}\|$ is used. This completes the proof of the lemma. ■

Appendix B
Proof of Theorem 17

Before proving Theorem 17, we first introduce two lemmas, whose proof can be found in [35].

Lemma 28 [35]: Suppose that $0 \le \boldsymbol{C}_k \le \boldsymbol{I}_p$, $k \ge 0$ with $\boldsymbol{C}_k \in \mathbb{R}^{p\times p}$. Set $\boldsymbol{\Pi}(k+1, j) = (\boldsymbol{I} - \boldsymbol{C}_k)\boldsymbol{\Pi}(k, j)$, $\boldsymbol{\Pi}(j, j) = \boldsymbol{I}$, $\forall\, k \ge j$, then we have

$$\|\boldsymbol{\Pi}(M, k)\| \le \left(1 - \frac{\lambda_{\min}(\boldsymbol{F}_{kM})}{2\left(1 + r_{kM}^2\right)}\right)^{\frac{1}{2}}, \quad M > k \qquad (36)$$

where $\boldsymbol{F}_{kM} \triangleq \sum_{j=k}^{M-1} \boldsymbol{C}_j$, $r_{kM} \triangleq \sum_{j=k}^{M-1} \|\boldsymbol{C}_j\|$, and $\lambda_{\min}(\boldsymbol{F}_{kM})$ represents the minimum eigenvalue of $\boldsymbol{F}_{kM}$.

Lemma 29 [35]: Suppose that $0 \le \boldsymbol{C}_k \le \boldsymbol{I}_p$, $k \ge 0$ with $\boldsymbol{C}_k \in \mathbb{R}^{p\times p}$. Then $\boldsymbol{\Pi}(k, 0)$ defined in Lemma 28 converges to 0 as $k \to \infty$, if there exists a sequence of monotonically increasing positive integers $\{t_k\}$ with $t_k \to \infty$ as $k \to \infty$, such that

$$\sum_{k=1}^{\infty} \frac{\lambda_{\min}(\boldsymbol{F}_k)}{1 + \lambda_{\max}^2(\boldsymbol{F}_k)} = \infty \qquad (37)$$

where $\boldsymbol{F}_k \triangleq \sum_{j=t_{k-1}}^{t_k-1} \boldsymbol{C}_j$.

Proof of Theorem 17:

Proof: Using Lemma 29, we just need to show that there exists an integer sequence $\{t_k\}$ with $t_k \to \infty$ as $k \to \infty$, such that

$$\sum_{k=1}^{\infty} \frac{\lambda_{\min}\left(\sum_{j=t_{k-1}}^{t_k-1} \mu \boldsymbol{G}_j\right)}{1 + \lambda_{\max}^2\left(\sum_{j=t_{k-1}}^{t_k-1} \mu \boldsymbol{G}_j\right)} = \infty. \qquad (38)$$

In the following, we will show this by three steps.

Step 1 (Construction of the Integer Sequence $\{t_k\}$): Without loss of generality, the constant K_0 in Assumption 6 can be taken to satisfy the inequality $\log \mathrm{tr}(\boldsymbol{R}_{K_0}) \ge 1$ since $\|\boldsymbol{R}_k\| \to \infty$ as $k \to \infty$. Thus, by the definition of d_k and $\hat{g}(t)$ in Lemma 15, we have for $t \ge 0$

$$t \le d_{\hat{g}(t)+1} \le d_{\hat{g}(t)} + 1 \le t + 1. \qquad (39)$$

We first prove the following result:

$$\hat{g}(t) \to \infty, \quad t \to \infty. \qquad (40)$$

By the boundedness of the condition number of $\boldsymbol{R}_k$, we have for large k

$$\begin{aligned}
\frac{\|\boldsymbol{R}_k\|}{\|\boldsymbol{R}_{k-1}\|} &= \frac{\max_i r_k^i}{\max_i r_{k-1}^i} = \frac{\max_i r_k^i}{\min_i r_k^i} \cdot \frac{\min_i r_k^i}{\max_i r_{k-1}^i} \\
&\le \frac{\max_i r_k^i}{\min_i r_k^i} \cdot \frac{r_k^{i_1}}{r_{k-1}^{i_1}} \le \gamma r^*.
\end{aligned}$$

Hence, we have

$$\frac{\mathrm{tr}\boldsymbol{R}_k}{\mathrm{tr}\boldsymbol{R}_{k-1}} \le \frac{n\|\boldsymbol{R}_k\|}{\|\boldsymbol{R}_{k-1}\|} \le n\gamma r^*. \qquad (41)$$

Then by (34) and (41), we have for any $k \ge K_0$

$$\begin{aligned}
d_k &= \sum_{j=K_0}^{k-1} \frac{\mathrm{tr}\boldsymbol{R}_j - \mathrm{tr}\boldsymbol{R}_{j-1}}{\mathrm{tr}(\boldsymbol{R}_j)(\log \mathrm{tr}(\boldsymbol{R}_{j-1}))^{\frac{1}{3}}} \\
&\ge \frac{1}{n\gamma r^*} \sum_{j=K_0}^{k-1} \frac{\mathrm{tr}\boldsymbol{R}_j - \mathrm{tr}\boldsymbol{R}_{j-1}}{\mathrm{tr}(\boldsymbol{R}_{j-1})(\log \mathrm{tr}(\boldsymbol{R}_{j-1}))^{\frac{1}{3}}} \\
&\ge \frac{1}{n\gamma r^*} \sum_{j=K_0}^{k-1} \int_{\mathrm{tr}\boldsymbol{R}_{j-1}}^{\mathrm{tr}\boldsymbol{R}_j} \frac{dx}{x(\log x)^{\frac{1}{3}}} \\
&= \frac{3}{2n\gamma r^*}\left(\log^{\frac{2}{3}} \mathrm{tr}\boldsymbol{R}_{k-1} - \log^{\frac{2}{3}} \mathrm{tr}\boldsymbol{R}_{K_0-1}\right). \qquad (42)
\end{aligned}$$

Since $\|\boldsymbol{R}_k\| \xrightarrow{k\to\infty} \infty$, then we have $d_k \to \infty$ as $k \to \infty$. By the definition of $\hat{g}(t)$, we further have $\hat{g}(t) < \infty$ for all $t > 0$ and $\hat{g}(t) \to \infty$ as $t \to \infty$.

By (40) and $\|\boldsymbol{R}_j\| \xrightarrow{j\to\infty} \infty$, we see that there exists sufficiently large N_1 such that for $j \geq \hat{g}(N_1)$

$$\frac{n(\log \mathrm{tr}\boldsymbol{R}_j)^{\frac{1}{3}}}{\mathrm{tr}\boldsymbol{R}_j} \leq \frac{1}{2N} \tag{43}$$

where N is defined in Assumption 6. The integer sequence $\{t_k\}$ can be taken as

$$t_k = \hat{g}(N_1 + k\alpha), \quad \alpha = 2Nm(n+2)+1. \tag{44}$$

Then by (40), we have $t_k \to \infty$ as $k \to \infty$.

Step 2 [Estimation of $\lambda_{\min}(\sum_{j=t_{k-1}}^{t_k-1} \mu \boldsymbol{G}_j)$]: By the properties of r_j^i, we have

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=t_{k-1}}^{t_k-1}\frac{\boldsymbol{\varphi}_j^i \boldsymbol{\varphi}_j^{i\,T}}{r_j^i} \\
&\quad\geq \sum_{i=1}^{n}\sum_{j=t_{k-1}}^{t_k}\frac{\boldsymbol{\varphi}_j^i \boldsymbol{\varphi}_j^{i\,T}}{r_j^i} - n\boldsymbol{I}_m \\
&\quad= \sum_{j=t_{k-1}}^{t_k}\sum_{i=1}^{n}\frac{1}{r_j^i}\left(\sum_{l=1}^{j}\boldsymbol{\varphi}_l^i\boldsymbol{\varphi}_l^{i\,T} - \sum_{l=1}^{j-1}\boldsymbol{\varphi}_l^i\boldsymbol{\varphi}_l^{i\,T}\right) - n\boldsymbol{I}_m \\
&\quad\geq \sum_{j=t_{k-1}}^{t_k}\frac{1}{\mathrm{tr}\boldsymbol{R}_j}\sum_{i=1}^{n}\left(\sum_{l=1}^{j}\boldsymbol{\varphi}_l^i\boldsymbol{\varphi}_l^{i\,T} - \sum_{l=1}^{j-1}\boldsymbol{\varphi}_l^i\boldsymbol{\varphi}_l^{i\,T}\right) - n\boldsymbol{I}_m \\
&\quad= \sum_{j=t_{k-1}+1}^{t_k+1}\frac{1}{\mathrm{tr}\boldsymbol{R}_{j-1}}\sum_{i=1}^{n}\sum_{l=1}^{j-1}\boldsymbol{\varphi}_l^i\boldsymbol{\varphi}_l^{i\,T} - n\boldsymbol{I}_m \\
&\qquad- \sum_{j=t_{k-1}}^{t_k}\frac{1}{\mathrm{tr}\boldsymbol{R}_j}\sum_{i=1}^{n}\sum_{l=1}^{j-1}\boldsymbol{\varphi}_l^i\boldsymbol{\varphi}_l^{i\,T}.
\end{aligned} \tag{45}$$

By the definition of $\lambda_{\min}^{(k)}$ in Assumption 6 and (45), we obtain

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=t_{k-1}}^{t_k-1}\frac{\boldsymbol{\varphi}_j^i \boldsymbol{\varphi}_j^{i\,T}}{r_j^i} \\
&\quad\geq \sum_{j=t_{k-1}+1}^{t_k}\left(\frac{1}{\mathrm{tr}\boldsymbol{R}_{j-1}} - \frac{1}{\mathrm{tr}\boldsymbol{R}_j}\right)\sum_{i=1}^{n}\sum_{l=1}^{j-1}\boldsymbol{\varphi}_l^i\boldsymbol{\varphi}_l^{i\,T} - n\boldsymbol{I}_m \\
&\qquad+ \frac{1}{\mathrm{tr}\boldsymbol{R}_{t_k}}\sum_{i=1}^{n}\sum_{l=1}^{t_k}\boldsymbol{\varphi}_l^i\boldsymbol{\varphi}_l^{i\,T} - \frac{1}{\mathrm{tr}\boldsymbol{R}_{t_{k-1}}}\sum_{i=1}^{n}\sum_{l=1}^{t_{k-1}}\boldsymbol{\varphi}_l^i\boldsymbol{\varphi}_l^{i\,T} \\
&\quad\geq \sum_{j=t_{k-1}+1}^{t_k}\left(\frac{\mathrm{tr}\left(\boldsymbol{\Phi}_j^T\boldsymbol{\Phi}_j\right)}{\mathrm{tr}\boldsymbol{R}_j\mathrm{tr}\boldsymbol{R}_{j-1}}\right)\sum_{i=1}^{n}\sum_{l=1}^{j-1}\boldsymbol{\varphi}_l^i\boldsymbol{\varphi}_l^{i\,T} - (n+1)\boldsymbol{I}_m \\
&\quad\geq \sum_{j=t_{k-1}+1}^{t_k}\left(\lambda_{\min}^{(j-1)} - \frac{n}{m}\right)\cdot\left(\frac{\mathrm{tr}\left(\boldsymbol{\Phi}_j^T\boldsymbol{\Phi}_j\right)}{\mathrm{tr}\boldsymbol{R}_j\mathrm{tr}\boldsymbol{R}_{j-1}}\right)\boldsymbol{I}_m \\
&\qquad-(n+1)\boldsymbol{I}_m.
\end{aligned} \tag{46}$$

Combining this with Assumption 6, we have

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=t_{k-1}}^{t_k-1}\frac{\boldsymbol{\varphi}_j^i \boldsymbol{\varphi}_j^{i\,T}}{r_j^i} + (n+1)\boldsymbol{I}_m \\
&\quad\geq \sum_{j=t_{k-1}+1}^{t_k}\left(\frac{\lambda_{\max}^{(j-1)}}{N(\log \mathrm{tr}\boldsymbol{R}_{j-1})^{\frac{1}{3}}} - \frac{n}{m}\right)\left(\frac{\mathrm{tr}\left(\boldsymbol{\Phi}_j^T\boldsymbol{\Phi}_j\right)}{\mathrm{tr}\boldsymbol{R}_j\mathrm{tr}\boldsymbol{R}_{j-1}}\right)\boldsymbol{I}_m \\
&\quad\geq \frac{1}{m}\sum_{j=t_{k-1}+1}^{t_k}\left(\frac{\mathrm{tr}\boldsymbol{R}_{j-1}}{N(\log \mathrm{tr}\boldsymbol{R}_{j-1})^{\frac{1}{3}}} - n\right)\left(\frac{\mathrm{tr}\left(\boldsymbol{\Phi}_j^T\boldsymbol{\Phi}_j\right)}{\mathrm{tr}\boldsymbol{R}_j\mathrm{tr}\boldsymbol{R}_{j-1}}\right)\boldsymbol{I}_m \\
&\quad= \frac{1}{m}\sum_{j=t_{k-1}+1}^{t_k}\left(\frac{1}{N} - \frac{n(\log \mathrm{tr}\boldsymbol{R}_{j-1})^{\frac{1}{3}}}{\mathrm{tr}\boldsymbol{R}_{j-1}}\right) \\
&\qquad\cdot\left(\frac{\mathrm{tr}\left(\boldsymbol{\Phi}_j^T\boldsymbol{\Phi}_j\right)}{\mathrm{tr}\boldsymbol{R}_j(\log \mathrm{tr}\boldsymbol{R}_{j-1})^{\frac{1}{3}}}\right)\boldsymbol{I}_m \\
&\quad\geq \frac{1}{2Nm}\sum_{j=t_{k-1}+1}^{t_k}\left(\frac{\mathrm{tr}\left(\boldsymbol{\Phi}_j^T\boldsymbol{\Phi}_j\right)}{\mathrm{tr}\boldsymbol{R}_j(\log \mathrm{tr}\boldsymbol{R}_{j-1})^{\frac{1}{3}}}\right)\boldsymbol{I}_m
\end{aligned}$$

where (43) is used in the last inequality. Hence, by the definition of d_k in Lemma 15, we can obtain

$$\sum_{i=1}^{n}\sum_{j=t_{k-1}}^{t_k-1}\frac{\boldsymbol{\varphi}_j^i \boldsymbol{\varphi}_j^{i\,T}}{r_j^i} \geq \frac{d_{t_k+1} - d_{t_{k-1}+1}}{2Nm}\boldsymbol{I}_m - (n+1)\boldsymbol{I}_m.$$

Furthermore, using Lemma 15, (39) and (44), we have

$$\begin{aligned}
&\lambda_{\min}\left(\sum_{j=t_{k-1}}^{t_k-1}\boldsymbol{G}_j\right) \\
&\quad\geq \sigma\,\lambda_{\min}\left(\sum_{i=1}^{n}\sum_{j=t_{k-1}}^{t_k-1}\frac{\boldsymbol{\varphi}_j^i \boldsymbol{\varphi}_j^{i\,T}}{r_j^i}\right) \\
&\quad\geq \frac{\sigma}{2Nm}(d_{\hat{g}(N_1+k\alpha)+1} - d_{\hat{g}(N_1+(k-1)\alpha+1)}) - \sigma(n+1) \\
&\quad\geq \frac{\sigma}{2Nm}(N_1 + k\alpha - (N_1 + (k-1)\alpha + 1)) - \sigma(n+1) \\
&\quad= \sigma\left(\frac{\alpha-1}{2Nm} - (n+1)\right) = \sigma.
\end{aligned} \tag{47}$$

Step 3 [Estimation of $\lambda_{\max}(\sum_{j=t_{k-1}}^{t_k-1} \mu \boldsymbol{G}_j)$]: By the basic properties of the trace and the Euclidean norm of the matrix, we have

$$\begin{aligned}
\sum_{j=t_{k-1}}^{t_k-1}\|\boldsymbol{A}_j\| &\leq \sum_{j=t_{k-1}}^{t_k-1}\mathrm{tr}\left(\boldsymbol{R}_j^{-1}\boldsymbol{\Phi}_j^T\boldsymbol{\Phi}_j\right) \leq \sum_{j=t_{k-1}}^{t_k-1}\mathrm{tr}\left(\boldsymbol{\Phi}_j^T\boldsymbol{\Phi}_j\right)\left\|\boldsymbol{R}_j^{-1}\right\| \\
&\leq \gamma\sum_{j=t_{k-1}}^{t_k-1}\frac{\mathrm{tr}\left(\boldsymbol{\Phi}_j^T\boldsymbol{\Phi}_j\right)}{\|\boldsymbol{R}_j\|} \leq n\gamma\sum_{j=t_{k-1}}^{t_k-1}\frac{\mathrm{tr}\left(\boldsymbol{\Phi}_j^T\boldsymbol{\Phi}_j\right)}{\mathrm{tr}\boldsymbol{R}_j} \\
&\leq n\gamma(\log \mathrm{tr}\boldsymbol{R}_{t_k-1})^{\frac{1}{3}}\sum_{j=t_{k-1}}^{t_k-1}\frac{\mathrm{tr}\left(\boldsymbol{\Phi}_j^T\boldsymbol{\Phi}_j\right)}{\mathrm{tr}\boldsymbol{R}_j(\log \mathrm{tr}\boldsymbol{R}_{j-1})^{\frac{1}{3}}}.
\end{aligned} \tag{48}$$

By the definition of $d_{\hat{g}(t)}$, (39), and (48), we have

$$\begin{aligned}
\sum_{j=t_{k-1}}^{t_k-1}\|\boldsymbol{A}_j\| &\leq n\gamma(\log \mathrm{tr}\boldsymbol{R}_{t_k-1})^{\frac{1}{3}}(d_{t_k} - d_{t_{k-1}}) \\
&= n\gamma(\log \mathrm{tr}\boldsymbol{R}_{t_k-1})^{\frac{1}{3}}(d_{\hat{g}(N_1+k\alpha)} - d_{\hat{g}(N_1+(k-1)\alpha)}) \\
&\leq n\gamma(\alpha+1)(\log \mathrm{tr}\boldsymbol{R}_{t_k-1})^{\frac{1}{3}}.
\end{aligned}$$

Then combining this with (9), we have

$$\begin{aligned}
\lambda_{\max}\left(\sum_{j=t_{k-1}}^{t_k-1}\boldsymbol{G}_j\right) &\leq \sum_{j=t_{k-1}}^{t_k-1}\|\boldsymbol{G}_j\| \leq (1+4\nu)\sum_{j=t_{k-1}}^{t_k-1}\|\boldsymbol{A}_j\| \\
&\leq n\gamma(1+4\nu)(\alpha+1)(\log \mathrm{tr}\boldsymbol{R}_{t_k-1})^{\frac{1}{3}}.
\end{aligned} \tag{49}$$

According to (39), (42) and (44), we can see that

$$N_1 + k\alpha \geq d_{\hat{g}(N_1+k\alpha)} \geq \frac{3}{2n\gamma r^*}\left(\log^{\frac{2}{3}} \mathrm{tr}\boldsymbol{R}_{t_k-1} - \log^{\frac{2}{3}} \mathrm{tr}\boldsymbol{R}_{K_0-1}\right). \quad (50)$$

Using (49) and (50), we have

$$\lambda_{\max}^2\left(\sum_{j=t_{k-1}}^{t_k-1} \boldsymbol{G}_j\right) \leq n^2\gamma^2(1+4\nu)^2(\alpha+1)^2 \cdot \left(\frac{2n\gamma r^*}{3}(N_1+k\alpha) + \log^{\frac{2}{3}} \mathrm{tr}\boldsymbol{R}_{K_0-1}\right) \triangleq p_k = O(k). \quad (51)$$

Combining this with (47) yields

$$\sum_{k=1}^{\infty} \frac{\lambda_{\min}\left(\sum_{j=t_{k-1}}^{t_k-1} \mu\boldsymbol{G}_j\right)}{1+\lambda_{\max}^2\left(\sum_{j=t_{k-1}}^{t_k-1} \mu\boldsymbol{G}_j\right)} \geq \sum_{k=1}^{\infty} \frac{\mu\sigma}{1+\mu^2 p_k} = \infty. \quad (52)$$

According to Lemma 29, we can see that $\boldsymbol{\Psi}(k,0) \to 0$ as $k \to \infty$. This completes the proof of the theorem. ■

Appendix C
Proof of Lemma 23

Proof: By the definition of $\boldsymbol{\Psi}(k, j)$ in (7), it is clear that we have 1).

Using Lemma 22, we have

$$\begin{aligned}
&\det \boldsymbol{\Psi}(k+1,0) \\
&= \prod_{j=0}^{k} \det(\boldsymbol{I}_{mn} - \mu\boldsymbol{G}_j) \geq \prod_{j=0}^{k} (\det(\boldsymbol{I}_{mn} - \boldsymbol{A}_j))^{\tau_1} \\
&= (\det(\boldsymbol{I}_{mn} - \boldsymbol{A}_0))^{\tau_1} \left(\prod_{i=1}^{n}\prod_{j=1}^{k} \frac{r_{j-1}^i}{r_j^i}\right)^{\tau_1} \\
&= (\det(\boldsymbol{I}_{mn} - \boldsymbol{A}_0))^{\tau_1} \prod_{i=1}^{n} \frac{1}{(r_k^i)^{\tau_1}} \\
&= \frac{1}{\det\left(\boldsymbol{R}_k^{\tau_1}\right)} \prod_{i=1}^{n} \left(1 - \left\|\boldsymbol{\varphi}_0^i\right\|^2\right)^{\tau_1} \\
&\geq \frac{1}{\|\boldsymbol{R}_k\|^{n\tau_1}} \prod_{i=1}^{n} \left(1 - \left\|\boldsymbol{\varphi}_0^i\right\|^2\right)^{\tau_1}. \quad (53)
\end{aligned}$$

Therefore, we have

$$\left(\frac{1}{\|\boldsymbol{R}_k\|^{n\tau_1}} \prod_{i=1}^{n} \left(1 - \left\|\boldsymbol{\varphi}_0^i\right\|^2\right)^{\tau_1}\right)^2 \leq \det(\boldsymbol{\Psi}(k+1,0)\boldsymbol{\Psi}^T(k+1,0)) \leq \|\boldsymbol{\Psi}(k+1,0)\|^{2mn}.$$

Since the initial value $\boldsymbol{\varphi}_0^i$ can be arbitrarily selected, without loss of generality we suppose $\|\boldsymbol{\varphi}_0^i\| \neq 1$ for all $i \in \{1, \ldots, n\}$. It is clear that 2) of the lemma holds.

Using Lemma 22, we have

$$\begin{aligned}
&\|\boldsymbol{\Psi}(k, j+1)\| \\
&\leq \|\boldsymbol{\Psi}(k,0)\|\|\boldsymbol{\Psi}^{-1}(j+1,0)\| \\
&\leq \|\boldsymbol{\Psi}(k,0)\| \prod_{p=1}^{j+1} \|(\boldsymbol{I}_{mn} - \mu\boldsymbol{G}_{p-1})^{-1}\| \\
&\leq \|\boldsymbol{\Psi}(k,0)\| \prod_{p=1}^{j+1} \det\left((\boldsymbol{I}_{mn} - \mu\boldsymbol{G}_{p-1})^{-1}\right) \\
&\leq \|\boldsymbol{\Psi}(k,0)\| \prod_{p=1}^{j+1} \frac{1}{\left(\det(\boldsymbol{I}_{mn} - \boldsymbol{A}_{p-1})\right)^{\tau_1}} \\
&= \|\boldsymbol{\Psi}(k,0)\| \left(\frac{1}{\det(\boldsymbol{I}_{mn} - \boldsymbol{A}_0)}\right)^{\tau_1} \prod_{i=1}^{n}\prod_{p=2}^{j+1} \left(\frac{r_{p-1}^i}{r_{p-2}^i}\right)^{\tau_1} \\
&= \|\boldsymbol{\Psi}(k,0)\| \cdot \left(\frac{1}{\det(\boldsymbol{I}_{mn} - \boldsymbol{A}_0)}\right)^{\tau_1} \left(\prod_{i=1}^{n} r_j^i\right)^{\tau_1} \\
&\leq \left(\frac{1}{\det(\boldsymbol{I}_{mn} - \boldsymbol{A}_0)}\right)^{\tau_1} \cdot \|\boldsymbol{\Psi}(k,0)\| \cdot \|\boldsymbol{R}_j\|^{n\tau_1}. \quad (54)
\end{aligned}$$

Hence, 3) of the lemma can be proved.

Now we will prove 4).

$$\begin{aligned}
\sum_{j=M+1}^{\infty} \frac{\|\boldsymbol{\Phi}_j\|^2}{\|\boldsymbol{R}_j\|^{1+\varsigma}} &\leq \sum_{j=M+1}^{\infty} \frac{\mathrm{tr}\left(\boldsymbol{\Phi}_j^T\boldsymbol{\Phi}_j\right)}{\frac{1}{n^{1+\varsigma}}(\mathrm{tr}\boldsymbol{R}_j)^{1+\varsigma}} \\
&= n^{1+\varsigma} \sum_{j=M+1}^{\infty} \int_{\mathrm{tr}\boldsymbol{R}_{j-1}}^{\mathrm{tr}\boldsymbol{R}_j} \frac{1}{(\mathrm{tr}\boldsymbol{R}_j)^{1+\varsigma}}\, dt \\
&\leq n^{1+\varsigma} \sum_{j=M+1}^{\infty} \int_{\mathrm{tr}\boldsymbol{R}_{j-1}}^{\mathrm{tr}\boldsymbol{R}_j} \frac{1}{t^{1+\varsigma}}\, dt \\
&\leq n^{1+\varsigma} \int_{\mathrm{tr}\boldsymbol{R}_M}^{\infty} \frac{1}{t^{1+\varsigma}}\, dt \\
&= \frac{n^{1+\varsigma}}{\varsigma} \frac{1}{(\mathrm{tr}\boldsymbol{R}_M)^{\varsigma}} \leq \frac{n^{1+\varsigma}}{\varsigma} \frac{1}{\|\boldsymbol{R}_M\|^{\varsigma}}.
\end{aligned}$$

This completes the proof of the lemma. ■

Appendix D
Proof of Theorem 25

Proof: Using (47) and (51) in Appendix B, there exists a constant $c^* > 0$ such that

$$\frac{\lambda_{\min}\left(\sum_{j=t_{k-1}}^{t_k-1} \mu\boldsymbol{G}_j\right)}{1+\lambda_{\max}^2\left(\sum_{j=t_{k-1}}^{t_k-1} \mu\boldsymbol{G}_j\right)} \geq \frac{c^*}{1+k}$$

where t_k is defined in (44) of Appendix B. In fact, since $\mu(1+4\nu) < 1$, we can take

$$c^* = \frac{\mu\sigma}{1+[n\gamma(\alpha+1)]^2 \max\left\{\frac{2n\gamma r^*\alpha}{3}, \frac{2n\gamma r^* N_1}{3} + \log^{\frac{2}{3}} \mathrm{tr}\boldsymbol{R}_{K_0-1}\right\}}$$

with $\alpha = 2Nm(n+2)+1$, $N_1 \propto n$, where $\sigma = ((l_{m+1}^2 a\nu)/(2n + l_{m+1}^2 a\nu n))$, and γ and r^* are, respectively, defined in Lemma 13 and Theorem 17.

Using Lemma 28, we have

$$\|\boldsymbol{\Psi}(t_k, t_{k-1})\| \leq \left(1 - \frac{\lambda_{\min}\left(\sum_{j=t_{k-1}}^{t_k-1} \mu\boldsymbol{G}_j\right)}{2\left(1 + \left(\sum_{j=t_{k-1}}^{t_k-1} \|\mu\boldsymbol{G}_j\|\right)^2\right)}\right)^{\frac{1}{2}}$$

$$\leq \left(1-\frac{\lambda_{\min}\left(\sum_{j=t_{k-1}}^{t_k-1}\mu \boldsymbol{G}_j\right)}{2\left(1+m^2n^2\lambda_{\max}^2\left(\sum_{j=t_{k-1}}^{t_k-1}\mu \boldsymbol{G}_j\right)\right)}\right)^{\frac{1}{2}}$$
$$\leq \left(1-\frac{\lambda_{\min}\left(\sum_{j=t_{k-1}}^{t_k-1}\mu \boldsymbol{G}_j\right)}{2m^2n^2\left(1+\lambda_{\max}^2\left(\sum_{j=t_{k-1}}^{t_k-1}\mu \boldsymbol{G}_j\right)\right)}\right)^{\frac{1}{2}}$$
$$\leq \left(1-\frac{\bar{c}}{1+k}\right)^{\frac{1}{2}} \quad \left(\bar{c}=\frac{c^*}{2m^2n^2}\right). \tag{55}$$

By the definition of t_k in (44), we have the following estimate about $\|\boldsymbol{\Psi}(\hat{g}(N_1+j\alpha),0)\|$ with N_1 and α defined in Appendix B

$$\begin{aligned}
&\|\boldsymbol{\Psi}(\hat{g}(N_1+j\alpha),0)\| \\
&\quad\leq \prod_{l=1}^{j}\|\boldsymbol{\Psi}(\hat{g}(N_1+l\alpha),\hat{g}(N_1+(l-1)\alpha)\| \\
&\qquad\cdot\|\boldsymbol{\Psi}(\hat{g}(N_1),0)\| \\
&\quad\leq \prod_{l=1}^{j}\left(1-\frac{\bar{c}}{1+l}\right)^{\frac{1}{2}} \leq \prod_{l=1}^{j}e^{-\frac{\bar{c}}{2(1+l)}} \\
&\quad= e^{-\sum_{l=1}^{j}\frac{\bar{c}}{2(1+l)}} \leq e^{-\frac{\bar{c}}{2}\cdot\log\left(\frac{j+2}{2}\right)} = \left(\frac{j+2}{2}\right)^{-\frac{\bar{c}}{2}}
\end{aligned} \tag{56}$$

where the inequalities $1-x \leq e^{-x}$ for all $x \geq 0$ and $\sum_{j=1}^{k}(1/(1+j)) \geq \log^{((k+2)/2)}$ for all $k \geq 1$ are used.

Since $\hat{g}(t) \to \infty$ as $t \to \infty$ in (40), then for any $k \geq \hat{g}(N_1+\alpha)$, there exists $j \geq 1$ such that

$$\hat{g}(N_1+j\alpha) \leq k \leq \hat{g}(N_1+(j+1)\alpha).$$

By the monotonicity of d_k and (39), we have

$$d_k \leq d_{\hat{g}(N_1+(j+1)\alpha)} \leq N_1+(j+1)\alpha.$$

Thus, $j \geq ((d_k-N_1-\alpha)/\alpha)$. According to (56), we obtain

$$\begin{aligned}
&\|\boldsymbol{\Psi}(k,0)\| \\
&\quad\leq \|\boldsymbol{\Psi}(k,\hat{g}(N_1+j\alpha))\|\cdot\|\boldsymbol{\Psi}(\hat{g}(N_1+j\alpha),0)\| \\
&\quad\leq \|\boldsymbol{\Psi}(\hat{g}(N_1+j\alpha),0)\| \leq \left(\frac{d_k-N_1+\alpha}{2\alpha}\right)^{-\frac{\bar{c}}{2}}.
\end{aligned} \tag{57}$$

Combining (42) with (57), we have for large k

$$\|\boldsymbol{\Psi}(k,0)\| = O(\log \mathrm{tr}\boldsymbol{R}_{k-1})^{-\frac{\bar{c}}{3}} = O(\log \mathrm{tr}\boldsymbol{R}_k)^{-\frac{\bar{c}}{3}}.$$

Hence using Lemma 24, we have $\|\widetilde{\boldsymbol{\Theta}}_k\| = O((\log\|\boldsymbol{R}_k\|)^{-\delta_1})$ with $\delta_1 = (\delta\bar{c}/(3n\tau_1(1+\delta))) = (\delta c^*/(6m^3n^4(1+\delta)))$. ■

REFERENCES

[1] Q. Y. Liu, Z. Wang, X. He, and D. H. Zhou, "On Kalman-consensus filtering with random link failures over sensor networks," *IEEE Trans. Autom. Control*, vol. 63, no. 8, pp. 2701–2708, Aug. 2018.

[2] D. Gan and Z. Liu, "Distributed order estimation of ARX model under cooperative excitation condition," *SIAM J. Control Optim.*, vol. 60, no. 3, pp. 1519–1545, Jun. 2022.

[3] D. Gan and Z. Liu, "Performance analysis of the compressed distributed least squares algorithm," *Syst. Control Lett.*, vol. 164, Jun. 2022, Art. no. 105228.

[4] A. H. Sayed, "Diffusion Adaptation over Networks," *Academic Press Library Signal Process.*, vol. 3, pp. 323–453, Jan. 2013.

[5] A. H. Sayed, S.-Y. Tu, J. Chen, X. Zhao, and Z. J. Towfic, "Diffusion strategies for adaptation and learning over networks: An examination of distributed strategies and network behavior," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 155–171, May 2013.

[6] N. Bogdanovic, J. Plata-Chaves, and K. Berberidis, "Distributed incremental-based LMS for node-specific adaptive parameter estimation," *IEEE Trans. Signal Process.*, vol. 62, no. 20, pp. 5382–5397, Oct. 2014.

[7] S. Battilotti, F. Cacace, M. d'Angelo, and A. Germani, "Asymptotically optimal consensus-based distributed filtering of continuous-time linear systems," *Automatica*, vol. 122, Dec. 2020, Art. no. 109189.

[8] S. Xie and L. Guo, "Analysis of distributed adaptive filters based on diffusion strategies over sensor networks," *IEEE Trans. Autom. Control*, vol. 63, no. 11, pp. 3643–3658, Nov. 2018.

[9] D. Gan, S. Xie, and Z. Liu, "Stability of the distributed Kalman filter using general random coefficients," *Sci. China Inf. Sci.*, vol. 64, no. 7, Jul. 2021, Art. no. 172204.

[10] S. S. Ram, V. V. Veeravalli, and A. Nedic, "Distributed and recursive parameter estimation in parametrized linear state-space models," *IEEE Trans. Autom. Control*, vol. 55, no. 2, pp. 488–492, Feb. 2010.

[11] S. S. Stankovic, M. S. Stankovic, and D. M. Stipanovic, "Decentralized parameter estimation by consensus based stochastic approximation," *IEEE Trans. Autom. Control*, vol. 56, no. 3, pp. 531–543, Mar. 2011.

[12] R. Carli, A. Chiuso, L. Schenato, and S. Zampieri, "Distributed Kalman filtering based on consensus strategies," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 4, pp. 622–633, May 2008.

[13] J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Trans. Autom. Control*, vol. AC-31, no. 9, pp. 803–812, Sep. 1986.

[14] J. Hu, L. Xie, and C. Zhang, "Diffusion Kalman filtering based on covariance intersection," *IEEE Trans. Signal Process.*, vol. 60, no. 2, pp. 891–902, Feb. 2012.

[15] I. D. Schizas, G. Mateos, and G. B. Giannakis, "Distributed LMS for consensus-based in-network adaptive processing," *IEEE Trans. Signal Process.*, vol. 57, no. 6, pp. 2365–2382, Jun. 2009.

[16] N. Takahashi, I. Yamada, and A. H. Sayed, "Diffusion least-mean squares with adaptive combiners: Formulation and performance analysis," *IEEE Trans. Signal Process.*, vol. 58, no. 9, pp. 4795–4810, Sep. 2010.

[17] F. S. Cattivelli and A. H. Sayed, "Diffusion LMS strategies for distributed estimation," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1035–1048, Mar. 2010.

[18] R. Arablouei, K. Doğançay, S. Werner, and Y.-F. Huang, "Adaptive distributed estimation based on recursive least-squares and partial diffusion," *IEEE Trans. Signal Process.*, vol. 62, no. 14, pp. 3510–3522, Jul. 2014.

[19] J. Lei and H.-F. Chen, "Distributed estimation for parameter in heterogeneous linear time-varying models with observations at network sensors," *Commun. Inf. Syst.*, vol. 15, no. 4, pp. 423–451, 2015.

[20] C. Chen, Z. Liu, and L. Guo, "Performance bounds of distributed adaptive filters with cooperative correlated signals," *Sci. China Inf. Sci.*, vol. 59, no. 11, Nov. 2016, Art. no. 112202.

[21] S. Xie and L. Guo, "Analysis of normalized least mean squares-based consensus adaptive filters under a general information condition," *SIAM J. Control Optim.*, vol. 56, no. 5, pp. 343–3404, 2018.

[22] Y. Bengio, L. Goodfellow, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[23] B. Swenson, A. Sridhar, and H. V. Poor, "On distributed stochastic gradient algorithms for global optimization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Barcelona, Spain, May 2020, pp. 8594–8598.

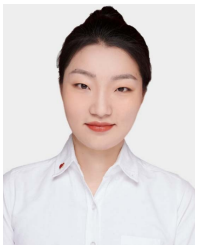
[24] D. Jakovetic, D. Bajovic, A. K. Sahu, and S. Kar, "Convergence rates for distributed stochastic optimization over random networks," in *Proc. IEEE Conf. Decis. Control (CDC)*, Miami Beach, FL, USA, Dec. 2018, pp. 4238–4245.

[25] M. Qi, W. Chen, Y. Wang, Z.-M. Ma, and T.-Y. Liu, "Convergence analysis of distributed stochastic gradient descent with shuffling," *Neurocomputing*, vol. 337, pp. 46–57, Apr. 2019.

[26] F. Barani, A. Savadi, and H. S. Yazdi, "Convergence behavior of diffusion stochastic gradient descent algorithm," *Signal Process.*, vol. 183, Jun. 2021, Art. no. 108014.

[27] Z. Li, B. Liu, and Z. Ding, "Consensus-based cooperative algorithms for training over distributed data sets using stochastic gradients," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 10, pp. 1–11, Oct. 2021, doi: 10.1109/TNNLS.2021.3071058.

[28] D. Yuan, D. W. C. Ho, and S. Xu, "Stochastic strongly convex optimization via distributed epoch stochastic gradient algorithm," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 6, pp. 2344–2357, Jun. 2021.
[29] S. Pu and A. Nedić, "Distributed stochastic gradient tracking methods," *Math. Program.*, vol. 187, pp. 409–457, May 2021.
[30] J. George, T. Yang, H. Bai, and P. Gurram, "Distributed stochastic gradient method for non-convex problems with applications in supervised learning," in *Proc. 58th IEEE Conf. Decis. Control (CDC)*, Dec. 2019, pp. 5538–5543.
[31] R. L. G. Cavalcante and S. Stanczak, "A distributed subgradient method for dynamic convex optimization problems under noisy information exchange," *IEEE J. Sel. Topics Signal Process.*, vol. 7, no. 2, pp. 243–256, Apr. 2013.
[32] J. B. Moore, "On strong consistency of least squares identification algorithms," *Automatica*, vol. 14, no. 5, pp. 505–509, Sep. 1978.
[33] W. Chen, C. Wen, S. Hua, and C. Sun, "Distributed cooperative adaptive identification and control for a group of continuous-time systems with a cooperative PE condition via consensus," *IEEE Trans. Autom. Control*, vol. 59, no. 1, pp. 91–106, Jan. 2014.
[34] H. Zhang, T. Wang, and Y. Zhao, "Asymptotically efficient recursive identification of FIR systems with binary-valued observations," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 51, no. 5, pp. 2687–2700, May 2021.
[35] L. Guo, *Time-Varying Stochastic Systems, Stability and Adaptive Theory*, 2nd ed. Beijing, China: Science Press, 2020.
[36] H. Chen and L. Guo, "Strong consistency of parameter estimates for discrete-time stochastic sytems," *J. Syst. Sci. Math. Sci.*, vol. 5, no. 2, pp. 81–93, 1985.
[37] R. Agaev and P. Chebotarev, "The matrix of maxium out forests of a digraph and its applications," *Autom. Remote Control*, vol. 61, no. 9, pp. 1424–1450, 2000.
[38] G. H. Hardy, J. E. Littlewood, and G. polya, *Inequalities*. Cambridge, U.K.: Cambridge Univ. Press, 1934.
[39] C. Hanfu and G. Lei, "Strong consistency of recursive identification by no use of persistent excitation condition," *Acta Mathematicae Applicatae Sinica*, vol. 2, no. 2, pp. 133–145, Jun. 1985.
[40] R. Olfati-Saber, "Distributed Kalman filtering for sensor networks," in *Proc. 46th IEEE Conf. Decis. Control*, New Orleans, LA, USA, Dec. 2007, pp. 5492–5498.
[41] G. Battistelli and L. Chisci, "Stability of consensus extended Kalman filter for distributed state estimation," *Automatica*, vol. 68, pp. 169–178, Jun. 2016.
[42] W. Li, Z. Wang, D. W. C. Ho, and G. Wei, "On boundedness of error covariances for Kalman consensus filtering problems," *IEEE Trans. Autom. Control*, vol. 65, no. 6, pp. 2654–2661, Jun. 2020.
[43] A. A. Ghandakly and A. M. Farhoud, "A parametrically optimized self-tuning regulator for power system stabilizers," *IEEE Trans. Power Syst.*, vol. 7, no. 3, pp. 1245–1250, Aug. 1992.
[44] K. J. Zachariah, J. W. Finch, and M. Farsi, "Multivariable self-tuning control of a turbine generator system," *IEEE Trans. Energy Convers.*, vol. 24, no. 2, pp. 406–414, Jun. 2009.
[45] H. Chen and P. Caines, "The strong consistency of the stochastic gradient algorithm of adaptive control," *IEEE Trans. Autom. Control*, vol. AC-30, no. 2, pp. 189–192, Feb. 1985.
[46] C. Godsil and G. Royle, *Algebraic Graph Theory*. New York, NY, USA: Springer-Verlag, 2001.
[47] L. Xiao, S. Boyd, and S. Lall, "A scheme for robust distributed sensor fusion based on average consensus," in *Proc. 4th Int. Symp. Inf. Process. Sensor Netw. (IPSN)*, Boise, ID, USA, Apr. 2005, pp. 63–70.



Die Gan received the B.S. degree in mathematics from Shandong University, Jinan, China, in 2017, and the Ph.D. degree in systems theory from the Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China, in 2022.

She is currently a Post-Doctoral Fellow with the Zhongguancun Laboratory, Beijing. Her research interests include system identification, machine learning, and distributed adaptive filter.

Dr. Gan received the China National Scholarship for graduate students in 2020 and the Outstanding Graduate Title of Beijing in 2022.

Zhixin Liu (Member, IEEE) received the B.S. degree in mathematics from Shandong University, Jinan, China, in 2002, and the Ph.D. degree in control theory from the Academy of Mathematics and Systems Science (AMSS), Chinese Academy of Sciences (CAS), Beijing, China, in 2007.

She had visiting positions with the KTH Royal Institute of Technology, Stockholm, Sweden, University of New South Wales, Kensington, Australia, and University of Maryland, College Park, MD, USA. She is currently a Professor with AMSS, CAS. She is a coauthor of the SIGEST Paper in SIREV in 2014. Her current research interests include complex systems and multiagent systems.

Prof. Liu is a co-recipient of the Best Theory Paper Award at the 13rd World Congress on Intelligent Control and Automation (WCICA).