2019 IEEE 58th Conference on Decision and Control (CDC)
Palais des Congrès et des Expositions Nice Acropolis
Nice, France, December 11-13, 2019

Strong consistency of the distributed stochastic gradient algorithm

Die Gan and Zhixin Liu

Abstract— **With the development of computer science and communication, the sensor networks are widely applied due to the advantages of flexibility, fault tolerance, and ease of deployment. In this paper, a distributed stochastic gradient (SG) algorithm is proposed where the distributed 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. The cooperative excitation condition is introduced, under which the strong consistency can be established for the distributed SG algorithm, without relying on the independency and stationarity assumptions of regression vectors which are commonly used in existing literature.**

I. INTRODUCTION

Filtering or parameter estimation is a very important problem in diverse fields including statistical learning, signal processing, system identification and adaptive control. With the development of computer science and communication, the sensor networks are widely applied due to the advantages of flexibility, fault tolerance, and ease of deployment. The sensor network brings more and more data, and how to apply the information from the sensors to design the proper estimation algorithm is a promising research direction.

Generally speaking, there are three methods to process the information from the sensors: centralized, distributed and a combination of both (cf., [1]). For the centralized method, the information measured by the sensors are transmitted to a fusion center, and the fusion center use all information to estimate the unknown signals or parameters. Compared with the distributed algorithms, the centralized ones are lack of robustness in addition to the burden brought by a large amount of computation and communication. In distributed algorithms, the sensors can cooperate to accomplish a complicated tasks in a cooperative manner even though each sensor can only receive local information, and has limited ability of computation and communication, The distributed estimation of filtering algorithms are widely applied in many practical engineering systems, such as target localization, noise elimination, see e.g., [2][3].

In the investigation of distributed estimation or filtering 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, i.e., incremental strategy [4], consensus strategy [5], and diffusion strategy [6]. Based on these three strategies, many different distributed estimation or filtering algorithms are proposed, such as the diffusion least mean squares, the consensus-based Kalman filter, the diffusion least squares. The stability and the convergence analysis of the distributed adaptive filtering and the distributed estimation algorithms are studied. Most results require that the regression signal satisfies independency and stationarity assumptions (cf., [7]-[11]). However, it is hard for the regression signals to satisfy the independency and stationarity assumptions because they are often produced by the feedback control systems, which makes it hard or even impossible to apply these theoretical results to practical systems. A preliminary attempt towards the relaxation of the independency and stationarity assumptions is made by Chen, Liu and Guo (cf., [12], [13]), where they provide a cooperative excitation condition to guarantee the stability of the diffusion least mean square algorithm. Recently, some elegant results for the distributed least mean square algorithms are established by Xie and Guo in [5][6] under a general cooperative information condition.

Compared with least mean square algorithm, the stochastic gradient algorithm has the advantages of simple expression and easy computation. In this paper, we focus on the investigation of the convergence properties of the distributed stochastic gradient algorithm. We first propose a distributed stochastic gradient algorithm by combining the consensus strategies and the diffusion of the regression vectors. We introduce a "weakest" excitation condition, under which the convergence of the algorithm can be established. Furthermore, we establish the convergence rate of the distributed stochastic gradient algorithm. It is worth mentioning that the work is based on [5][14][15][16][17], the properties of the product of stochastic matrices are obtained, which plays key role for our analysis. Our results are obtained without relying on the assumptions of the independency and stationarity assumptions.

The rest of this paper is organized as follows. We first introduce some notations and preliminaries on the distributed stochastic gradient algorithm In Section II. The strong consistency of the proposed algorithm is established in Section III and then the convergence rate is given in Section IV. The concluding remarks are made in the last section.

II. PROBLEM FORMULATION

A. Some Preliminaries

In this paper, 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 the

This work was supported by the National Key R&D Program of China under Grant 2018YFA0703800.

D. Gan and Z. X. Liu are with the Key Laboratory of Systems and Control, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, and School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, P. R. China. Emails: gandie@amss.ac.cn, lzx@amss.ac.cn

spectral norm induced by the Euclidean norm, i.e., $\|\boldsymbol{A}\| \triangleq (\lambda_{max}(\boldsymbol{A}\boldsymbol{A}^T))^{\frac{1}{2}}$, where $(\cdot)^T$ denotes the transpose operator and $\lambda_{\max}\{\cdot\}$ denotes the largest eigenvalue of the matrix. The notations $\det(\cdot)$ and $tr(\cdot)$ are used to denote the determinant and trace of the corresponding matrix respectively. If all 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

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{pmatrix} a_{11}\boldsymbol{B} & \cdots & a_{1n}\boldsymbol{B} \\ \vdots & \ddots & \vdots \\ a_{n1}\boldsymbol{B} & \cdots & a_{nn}\boldsymbol{B} \end{pmatrix} \in \mathbb{R}^{mp\times nq}.$$

An undirected weighted graph $\mathscr{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$ is used to describe the relationship between sensors, where $\mathcal{V} = \{1, 2, 3, \cdots, n\}$ is the set of sensors, the edge set $\mathcal{E} \in \mathcal{V} \times \mathcal{V}$ 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 and only if $(i, j) \in \mathcal{E}$. The neighbor set of the sensor i is denoted as $N_i = \{j \in \mathcal{V}, (i, j) \in \mathcal{E}\}$, and each sensor can only communicate with its neighbors. A path of length ℓ is a sequence of nodes $\{i_1, ...i_\ell\}$ satisfying $(i_j, i_{j+1}) \in \mathcal{E}$ for all $1 \le j \le \ell - 1$. The graph $\mathscr{G}$ is called connected if for any two sensors i and j, there is a path connecting them, and the diameter $D(\mathscr{G})$ of the graph $\mathscr{G}$ denotes the maximum length of the path between any two sensors. In this paper, we consider the convergence property of the distributed SG algorithm under the condition that the weighted matrix is symmetric and stochastic. Hence, the Laplacian matrix $\boldsymbol{L}$ of the graph $\mathscr{G}$ can be written as $\boldsymbol{L} = \boldsymbol{I} - \mathcal{A}$ with $\boldsymbol{I}$ being the identity matrix.

A classical result for the Laplacian matrix $\boldsymbol{L}$ can be stated as follows.

Lemma 1 [18] The Laplacian matrix $\boldsymbol{L}$ has at least one zero eigenvalue, with other eigenvalues positive and less than or equal to 2. Moreover, if the graph $\mathscr{G}$ is connected, then $\boldsymbol{L}$ has only one zero eigenvalue.

B. Distributed SG Algorithm

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

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

where y_k^i is the scalar observation of the sensor i at time k, $\{\varepsilon_k^i\}$ is a random noise process, φ_k^i is an m-dimension regression vector of i and usually contains input and output information, $\boldsymbol{\theta}$ is an unknown m-dimensional parameter.

The purpose of this paper is to propose a distributed algorithm to estimate the unknown parameter $\boldsymbol{\theta}$. For this, we propose the distributed SG algorithm which combines the consensus strategy of the estimation of neighbors with the diffusion of regression vectors. The detailed algorithm can be found in Table I.

TABLE I
DISTRIBUTED SG ALGORITHM

Given any initial estimates $\hat{\boldsymbol{\theta}}_0^i$ of each sensor i, the distributed SG algorithm is given as follows:

Step 1. For any $i \in \{1, ..., n\}$, set the initial value at each time k :
$$x_k^i(0) = \frac{\|\boldsymbol{\varphi}_k^i\|^2}{r_k^i}.$$

Step 2. Perform the following diffusion process for Q steps:
$$x_k^i(q+1) = \sum_{j\in N^i} a_{ij} x_k^j(q).$$
where $Q \ge D(\mathscr{G})$ with $D(\mathscr{G})$ being the diameter of $\mathscr{G}$.

Step 3. After Step 2, update the estimates of each sensor as follow:
$$\boldsymbol{z}_k^i = x_k^i(Q) \sum_{l\in N^i} a_{li}(\hat{\boldsymbol{\theta}}_k^i - \hat{\boldsymbol{\theta}}_k^l),$$
$$\hat{\boldsymbol{\theta}}_{k+1}^i = \hat{\boldsymbol{\theta}}_k^i + \mu \frac{\boldsymbol{\varphi}_k^i}{r_k^i}(y_{k+1}^i - (\boldsymbol{\varphi}_k^i)^T \hat{\boldsymbol{\theta}}_k^i) - \mu\nu \sum_{j\in N^i} a_{ij} \left(\boldsymbol{z}_k^i - \boldsymbol{z}_k^j\right),$$
$$r_k^i = 1 + \sum_{j=1}^{k} \left\|\boldsymbol{\varphi}_j^i\right\|^2.$$

For convenience of analysis, we introduce the following notations, see Table II.

TABLE II
SOME NOTATIONS

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

In order to proceed our theoretical analysis of the convergence property of the distributed SG algorithm, we introduce the following assumptions concerning the graph, regression vectors and the system noise.

Assumption 1 *The graph $\mathscr{G}$ is connected and contains self-loops at each sensor i, and the weighted matrix $\mathcal{A}$ is symmetric and stochastic.*

Assumption 2 *(Cooperative Excitation Condition) There exist two positive constants N and N_0 such that for $k \ge N_0$, the following inequality is satisfied,*

$$\frac{\lambda_{\max}^k}{\lambda_{\min}^k} \le N \left(log(tr\boldsymbol{R}_k)\right)^{\frac{1}{3}}, \tag{2}$$

where λ_{max}^k, λ_{min}^k represent the maximum and minimum eigenvalues of $\frac{n}{m}\boldsymbol{I}_m + \sum_{i=1}^{n}\sum_{j=1}^{k} \boldsymbol{\varphi}_j^i \boldsymbol{\varphi}_j^{i^T}$.

Without loss of generality, the constant N_0 in Assumption 2 can be taken to satisfy $\log tr(\boldsymbol{R}_{N_0}) \ge 1$.

Remark 1 *In [14], Guo proved that under the following excitation condition*

$$\lambda_{\max}\left\{\sum_{j=1}^{k}\boldsymbol{\varphi}_j^i\boldsymbol{\varphi}_j^{i\,T}\right\}/\lambda_{\min}\left\{\sum_{j=1}^{k}\boldsymbol{\varphi}_j^i\boldsymbol{\varphi}_j^{i\,T}\right\}\leq N\left(\log r_k^i\right)^{\frac{1}{3}}, \tag{3}$$

the convergence of the standard SG algorithm can be guaranteed, which can be regarded as the "weakest" excitation condition in the current literature. Assumption 2 *is introduced based on the condition (3), and can be considered as an extension to the distributed algorithm.*

Assumption 3 *We assume that the system noise is a martingale difference sequence, that is,* $E(\boldsymbol{\Xi}_{k+1}|\mathscr{F}_k)=0$ *with* $\mathscr{F}_k=\sigma\{\boldsymbol{\varphi}_j^i,\varepsilon_j^i,i=1,...n,j\leq k\}$ *and* $E(\cdot|\mathscr{F}_k)$ *being the conditional mathematical expectation, and there exist constants* $c_0>0$ *and* $\varepsilon\in[0,1)$ *(which may depend on* ω*) such that* $E(\|\boldsymbol{\Xi}_{k+1}\|^2|\mathscr{F}_k)\leq c_0\|\boldsymbol{R}_k\|^{\varepsilon}$.

Remark 2 *It is clear that if* $E(\|\boldsymbol{\Xi}_{k+1}\|^2|\mathscr{F}_k)\leq c_0$ *holds for all* k*, then we have* $E(\|\boldsymbol{\Xi}_{k+1}\|^2|\mathscr{F}_k)\leq c_0\|\boldsymbol{R}_k\|^{\varepsilon}$.

By (1), we can rewrite the distributed SG algorithm in Table I into the following matrix form,

$$\begin{aligned}\boldsymbol{Y}_{k+1}&=\boldsymbol{\Theta}^T\boldsymbol{\Phi}_k+\boldsymbol{\Xi}_{k+1},\\ \hat{\boldsymbol{\Theta}}_{k+1}&=\hat{\boldsymbol{\Theta}}_k+\mu\boldsymbol{\Phi}_k\boldsymbol{R}_k^{-1}(\boldsymbol{Y}_{k+1}^T-\boldsymbol{\Phi}_k^T\hat{\boldsymbol{\Theta}}_k)\\ &\quad-\mu\nu\mathscr{L}(\boldsymbol{X}_k(Q)\otimes\boldsymbol{I}_m)\mathscr{L}\hat{\boldsymbol{\Theta}}_k.\end{aligned} \tag{4}$$

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

$$\widetilde{\boldsymbol{\Theta}}_{k+1}=(I-\mu\boldsymbol{G}_k)\widetilde{\boldsymbol{\Theta}}_k-\mu\boldsymbol{\Phi}_k\boldsymbol{R}_k^{-1}\boldsymbol{\Xi}_{k+1}^T. \tag{5}$$

III. Strong Consistency of Parameter Estimates

Let matrix $\Phi(k,j)$ be recursively defined by

$$\boldsymbol{\Phi}(k+1,j)=(\boldsymbol{I}_{mn}-\mu\boldsymbol{G}_k)\boldsymbol{\Phi}(k,j),\quad\boldsymbol{\Phi}(j,j)=\boldsymbol{I}_{mn}. \tag{6}$$

In this section, we will establish a necessary and sufficient condition for strong consistency of the proposed distributed SG algorithm. For this purpose, we first list some lemmas whose proofs are omitted due to space limitations.

Lemma 2 *Suppose that Assumption* 1 *is satisfied, If* $\mu>0,\nu>0$ *and* $\mu(1+4n\nu)\leq 1$*, then we have*

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

Lemma 3 [19] Let $D_t\triangleq 1+\sum_{j=1}^{t}d_j,\ d_j\geq 0$, then

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

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

The following lemma provides the upper bound of the cumulative summation of the noises, which is an important step towards the convergence analysis of the algorithm.

Lemma 4 *Suppose that Assumption* 3 *is satisfied, the condition number of* $\boldsymbol{R}_k$ *is bounded(i.e there exits a positive constant* γ *which may depend on the sample* ω *such that* $\max_i r_k^i/\min_i r_k^i\leq\gamma$*), then* $\sum_{j=0}^{k}\boldsymbol{\Phi}_j\boldsymbol{R}_j^{-1}\boldsymbol{\Xi}_{j+1}^T$ *tends to a finite limit* $\boldsymbol{S}$ *as* $k\to\infty$*. Furthermore, there exit two positive constants* c *and* δ *which may depend upon the sample* ω *such that*

$$\left\|\boldsymbol{S}-\sum_{j=0}^{k-1}\boldsymbol{\Phi}_j\boldsymbol{R}_j^{-1}\boldsymbol{\Xi}_{j+1}^T\right\|\leq c\|\boldsymbol{R}_k\|^{-\delta}. \tag{7}$$

Lemma 5 *Assume that the steps* μ *and* ν *satisfy* $\mu(1+4n\nu)\leq 1$*, then for any* $k\geq 0$ *the following inequality holds,*

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

where $\boldsymbol{B}_j^2=\mu\boldsymbol{G}_j$.

Proof. By (6) and Lemma 2, we have

$$\begin{aligned}mn&=tr\boldsymbol{\Phi}(k,k)\boldsymbol{\Phi}^T(k,k)\\ &\geq tr\sum_{j=0}^{k-1}\Big[\boldsymbol{\Phi}(k,j+1)\boldsymbol{\Phi}^T(k,j+1)\\ &\quad-\boldsymbol{\Phi}(k,j)\boldsymbol{\Phi}^T(k,j)\Big]\\ &=tr\sum_{j=0}^{k-1}\Big\{\boldsymbol{\Phi}(k,j+1)\Big[\boldsymbol{I}_{mn}-\boldsymbol{\Phi}(j+1,j)\\ &\quad\cdot\boldsymbol{\Phi}^T(j+1,j)\Big]\boldsymbol{\Phi}^T(k,j+1)\Big\}\\ &\geq tr\sum_{j=0}^{k-1}\boldsymbol{\Phi}(k,j+1)\mu\boldsymbol{G}_j\boldsymbol{\Phi}^T(k,j+1)\\ &\geq\sum_{j=0}^{k-1}\|\boldsymbol{\Phi}(k,j+1)\boldsymbol{B}_j\|^2,\end{aligned}$$

which completes the proof. ∎

Now, we present the first theorem of the convergence of the distributed SG algorithm.

Theorem 1 *Suppose that the condition number of* $\boldsymbol{R}_k$ *is bounded, and* $\mu(1+4n\nu)<1$*. Then under Assumptions* 1 *and* 3*, for any initial value* $\hat{\boldsymbol{\Theta}}_0$ *the estimate* $\hat{\boldsymbol{\Theta}}_k$ *defined by* (4) *converges to the true parameter* $\boldsymbol{\Theta}$ *if and only if* $\boldsymbol{\Phi}(k,0)\to 0,\ k\to\infty$.

Proof. By (5) and (6), we have the following expression

$$\begin{aligned}\widetilde{\boldsymbol{\Theta}}_{k+1}&=\boldsymbol{\Phi}(k+1,0)\widetilde{\boldsymbol{\Theta}}_0\\ &\quad-\mu\sum_{j=0}^{k}\boldsymbol{\Phi}(k+1,j+1)\boldsymbol{\Phi}_j\boldsymbol{R}_j^{-1}\boldsymbol{\Xi}_{j+1}^T.\end{aligned} \tag{8}$$

We first prove the necessity part of the theorem. For any $\widetilde{\boldsymbol{\Theta}}_0$, we have $\widetilde{\boldsymbol{\Theta}}_k\to 0$. Note that the second term on the right-hand side of (8) is independent of $\widetilde{\boldsymbol{\Theta}}_0$. Thus, for

any $\widetilde{\boldsymbol{\Theta}}_0$, we have $\boldsymbol{\Phi}(k+1,0)\widetilde{\boldsymbol{\Theta}}_0 \to 0$ as $k \to \infty$, which means that $\boldsymbol{\Phi}(k+1,0) \to 0$ as $k \to \infty$.

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

$$\sum_{j=0}^{k} \boldsymbol{\Phi}(k+1,j+1)\boldsymbol{\Phi}_j \boldsymbol{R}_j^{-1}\boldsymbol{\Xi}_{j+1}^T \to 0,\ k \to \infty. \quad (9)$$

Set

$$\boldsymbol{S}_k = \sum_{j=1}^{k} \boldsymbol{\Phi}_j \boldsymbol{R}_j^{-1}\boldsymbol{\Xi}_{j+1}^T, \widetilde{\boldsymbol{S}}_k = \sum_{j=k+1}^{\infty} \boldsymbol{\Phi}_j \boldsymbol{R}_j^{-1}\boldsymbol{\Xi}_{j+1}^T,$$
$$\boldsymbol{S}_{-1} = 0.$$

By Lemma 4 we have $\|\widetilde{\boldsymbol{S}}_{k-1}\| \le c\|\boldsymbol{R}_k\|^{-\delta}$. Then

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

By Hölder inequality, the last term of the right hand side of the above equation can be estimated according to the following manner,

$$\begin{aligned}
&\|\sum_{j=0}^{k} \boldsymbol{\Phi}(k+1,j+1)[\boldsymbol{I}_{mn} - \boldsymbol{\Phi}(j+1,j)]\widetilde{\boldsymbol{S}}_{j-1}\| \\
&\le \sum_{j=0}^{k} \|\boldsymbol{\Phi}(k+1,j+1)\boldsymbol{B}_j\| \frac{\|\boldsymbol{B}_j\|}{\|\boldsymbol{R}_j\|^{\delta}} \\
&+ \left(\sum_{j=M+1}^{k} \|\boldsymbol{\Phi}(k+1,j+1)\boldsymbol{B}_j\|^2\right)^{\frac{1}{2}} \\
&\quad \cdot \left(\sum_{j=M+1}^{k} \frac{\|\boldsymbol{B}_j\|^2}{\|\boldsymbol{R}_j\|^{2\delta}}\right)^{\frac{1}{2}}. \quad (10)
\end{aligned}$$

By Lemmas 3 and 5 we have

$$\begin{aligned}
\sum_{j=1}^{\infty} \frac{\|\boldsymbol{B}_j\|^2}{\|\boldsymbol{R}_j\|^{2\delta}} &= \sum_{j=1}^{\infty} \frac{\|\boldsymbol{G}_j\|}{\|\boldsymbol{R}_j\|^{2\delta}} \\
&\le (1+4mn\nu)\sum_{j=1}^{\infty} \frac{\|\boldsymbol{A}_j\|}{\|\boldsymbol{R}_j\|^{2\delta}} \\
&\le (1+4mn\nu)\sum_{j=1}^{\infty} \frac{\|\boldsymbol{\Phi}_j\|^2\|\boldsymbol{R}_j^{-1}\|}{\|\boldsymbol{R}_j\|^{2\delta}} \\
&\le (1+4mn\nu)\gamma\sum_{j=1}^{\infty} \frac{\|\boldsymbol{\Phi}_j\|^2}{\|\boldsymbol{R}_j\|^{1+2\delta}} < \infty. \quad (11)
\end{aligned}$$

According to (11) and Lemma 5, $\|\sum_{j=0}^{k} \boldsymbol{\Phi}(k+1,j+1)\boldsymbol{\Phi}_j \boldsymbol{R}_j^{-1}\boldsymbol{\Xi}_{j+1}^T\|$ tends to zero if we first let $k \to \infty$, and then let $M \to \infty$. Hence (9) holds. This completes the proof of the theorem. ■

A key problem still remains unresolved: what conditions on the regression signals can guarantee that $\boldsymbol{\Phi}(k,0) \to 0$ as $k \to \infty$? In the following, we will prove that under the cooperative excitation condition (i.e., Assumption 2) we can establish the convergence results of the distributed algorithm.

We obtain the following results.

Theorem 2 *Let $\mu(1+4n\nu) \le 1$. Suppose that there exist $i_1, i_2 \in \{1,...n\}$ such that $\limsup_{k\to\infty} \frac{r_k^{i_1}}{r_{k-1}^{i_1}} < \infty$, $r_k^{i_2} \to \infty$, and the condition number of $\boldsymbol{R}_k$ is bounded. Under Assumptions 1 and 2, we have $\boldsymbol{\Phi}(k,0) \to 0,\ k \to \infty$.*

The proof of the theorem is very complicated, and we omit it due to space limitations.

Remark 3 *The above theorem shows that under the cooperative excitation condition, the convergence of the distributed SG algorithm can be established. Different form 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 proposed algorithm.

Lemma 6 *If $\mu(1+4n\nu) < 1$, then there exists a constant $\tau_1 > 1$ 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 $\boldsymbol{G}_k$, we have

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

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

Lemma 7 *If $\mu(1+4n\nu) < 1$, then we have the following inequalities,*

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

(ii) $\frac{1}{\|\boldsymbol{R}_k\|^{\tau_1}} = O(\|\boldsymbol{\Phi}(k+1,0)\|^{2m})\ , \quad \forall k \ge 1\ ;$

(iii) $\|\boldsymbol{\Phi}(k,j+1)\| = O(\|\boldsymbol{\Phi}(k,0)\| \cdot \|\boldsymbol{R}_j\|^{n\tau_1}), \forall k \geq j;$

(iv) $\sum_{j=M+1}^{\infty} \frac{\|\boldsymbol{\Phi}_j\|^2}{\|\boldsymbol{R}_j\|^{1+\delta}} \leq \frac{n^{1+\delta}}{\delta} \frac{1}{\|\boldsymbol{R}_M\|^{\delta}}$, $\forall k \geq 0.$

The proof of this Lemma is based on Lemma 6.

Theorem 3 *Under the conditions of Theorem 1, then we have*

$$\|\hat{\widetilde{\boldsymbol{\Theta}}}_k - \boldsymbol{\Theta}\| = O(\|\boldsymbol{\Phi}(k,0)\|^{\frac{\delta}{n\tau_1(1+\delta)}}) \quad a.s. \quad as \quad k \to \infty$$

Proof. Let

$$\alpha(t) = max\{j : \|\boldsymbol{R}_j\|^{n\tau_1} \leq t\}, \qquad t \geq 0,$$
$$\lambda(k) = \alpha(\|\boldsymbol{\Phi}(k,0)\|^{-\frac{1}{1+\delta}}), \quad k \geq 0.$$

By the definition of $\alpha(t)$ and $\lambda(k)$, we have $\|\boldsymbol{R}_{\alpha(t)}\|^{n\tau_1} \leq t$, and hence $\|\boldsymbol{R}_{\lambda(k)}\|^{n\tau_1} \leq \|\boldsymbol{\Phi}(k,0)\|^{-\frac{1}{1+\delta}}$.
According to Lemma 7 (iii), we have

$$\begin{aligned}\|\boldsymbol{\Phi}(k,\lambda(k)+1)\| &= O(\|\boldsymbol{\Phi}(k,0)\| \cdot \|\boldsymbol{R}_{\lambda(k)}\|^{n\tau_1}) \\ &= O(\boldsymbol{\Phi}(k,0)^{\frac{\delta}{1+\delta}}).\end{aligned}$$

Thus for large k, we have $\lambda(k) < k-1$. By Theorem 1, then we have the following estimation on the noise of the system,

$$\begin{aligned}&\|\sum_{j=0}^{k-1} \boldsymbol{\Phi}(k,j+1)\boldsymbol{\Phi}_j \boldsymbol{R}_j^{-1} \boldsymbol{\Xi}_{j+1}^T\| \\ &\leq \|\widetilde{\boldsymbol{S}}_{k-1}\| + \|\boldsymbol{\Phi}(k,0)\boldsymbol{S}\| + c\sum_{j=0}^{k-1} \|\boldsymbol{\Phi}(k,j+1)\| \cdot \frac{\|\mu \boldsymbol{G}_j\|}{\|\boldsymbol{R}_j\|^{\delta}} \\ &= O(\|\boldsymbol{R}_{\lambda(k)+1}\|^{-\delta}) + O(\|\boldsymbol{\Phi}(k,0)\|) \\ &\quad + O(\sum_{j=0}^{k-1} \|\boldsymbol{\Phi}(k,j+1)\| \cdot \frac{\|\boldsymbol{\Phi}_j\|^2}{\|\boldsymbol{R}_j\|^{1+\delta}}). \qquad (12)\end{aligned}$$

Now, we are in a position to estimate the last term of the right hand side of the above inequality. By Lemma 7, we have

$$\begin{aligned}&\sum_{j=0}^{k-1} \|\boldsymbol{\Phi}(k,j+1)\| \cdot \frac{\|\boldsymbol{\Phi}_j\|^2}{\|\boldsymbol{R}_j\|^{1+\delta}} \\ &\leq \sum_{j=0}^{\lambda(k)} \|\boldsymbol{\Phi}(k,\lambda(k)+1)\| \cdot \|\boldsymbol{\Phi}(\lambda(k)+1,j+1)\| \cdot \frac{\|\boldsymbol{\Phi}_j\|^2}{\|\boldsymbol{R}_j\|^{1+\delta}} \\ &\quad + \sum_{j=\lambda(k)+1}^{k-1} \|\boldsymbol{\Phi}(k,j+1)\| \cdot \frac{\|\boldsymbol{\Phi}_j\|^2}{\|\boldsymbol{R}_j\|^{1+\delta}} \\ &= O(\|\boldsymbol{\Phi}(k,0)\|^{\frac{\delta}{1+\delta}}) \sum_{j=0}^{\infty} \frac{\|\boldsymbol{\Phi}_j\|^2}{\|\boldsymbol{R}_j\|^{1+\delta}} + \sum_{j=\lambda(k)+1}^{k-1} \frac{\|\boldsymbol{\Phi}_j\|^2}{\|\boldsymbol{R}_j\|^{1+\delta}} \\ &\leq O(\|\boldsymbol{\Phi}(k,0)\|^{\frac{\delta}{1+\delta}}) + \frac{\|\boldsymbol{\Phi}_{\lambda(k)+1}\|^2}{\|\boldsymbol{R}_{\lambda(k)+1}\|^{1+\delta}} + \sum_{j=\lambda(k)+2}^{\infty} \frac{\|\boldsymbol{\Phi}_j\|^2}{\|\boldsymbol{R}_j\|^{1+\delta}} \\ &= O(\|\boldsymbol{\Phi}(k,0)\|^{\frac{\delta}{n\tau_1(1+\delta)}}). \qquad (13)\end{aligned}$$

Combining the above inequality with (8) and (12), then we have

$$\widetilde{\boldsymbol{\Theta}}_k = O(\|\boldsymbol{\Phi}(k,0)\|^{\frac{\delta}{n\tau_1(1+\delta)}}) \quad a.s. \quad as \quad k \to \infty, \qquad (14)$$

where $\|\boldsymbol{\Phi}(k,0)\| \leq 1$ is used in the above inequality. This completes the proof of the theorem. ■

Theorem 4 *Under the conditions of Theorem 3, If Assumption 2 is also satisfied, and there exists* $i_1 \in \{1,...,n\}$*, such that* $\limsup_{k\to\infty} \frac{r_k^{i_1}}{r_{k-1}^{i_1}} < \infty$*, then*

$$\widetilde{\boldsymbol{\Theta}}_k = O((\log\|\boldsymbol{R}_k\|)^{-\delta_1}) \quad \delta_1 > 0 \quad a.s. \quad as \quad k \to \infty \quad (15)$$

V. CONCLUDING REMARKS

In order to cooperatively estimate an unknown time-invariant parameter, we proposed a distributed SG algorithm based on the consensus strategies and the diffusion of the regression vectors. We introduced the cooperative excitation condition, under which the convergence of the proposed algorithm can be proved, and the convergence rate of the algorithm can be established. Moreover, we extend the convergence results to the case with correlated noise. In comparison with the existing results in the literature, our results are obtained without relying on the independency and stationarity assumptions. Many interesting problems deserve to be further investigated, for example, how to remove the boundedness of the condition number of $\boldsymbol{R}_k$, the analysis of other distributed algorithms such as distributed Kalman filter and distributed forgetting factor least square algorithm, and the combination of the distributed estimation with the distributed control.

REFERENCES

[1] Y. Bar-Shalom, Multitarget-Multisensor Tracking: Advanced Applications, *Norwood,* MA, Artech House, 1990.

[2] A. H. Sayed, Diffusion adaptation over networks, vol. 3, pp. 323-453, 2014.

[3] A. H. Sayed, S. Y. Tu, J. Chen, et al, Diffusion strategies for adaption and learning over networks: an examination of distributed strategies and network behavior[J], *IEEE Signal Processing Maganize*, vol. 30, no. 3, pp. 155-171, 2013.

[4] C. G. Lopes, A. H. Sayed, Incremental adaptive strategies over distributed networks, *IEEE Transactions Sinal Processing*, vol. 55, no. 8, pp. 4064-4077, 2007.

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

[6] S. Y. Xie, L. Guo, Analysis of distributed adaptive filters based on diffusion strategies over sensor networks, *IEEE Transactions on Automatic Control*, vol. 63, no. 11, pp. 3643-3648, 2018.

[7] S. Haykin, Adaptive filter theory, *Prentice-Hall*, Inc., 1986.

[8] F. S. Cattivelli, and A. H. Sayed, Diffusion LMS strategies for distributed estimation, *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1035-1048, 2010.

[9] I. D. Schizas, G. Mateos, and G. B. Giannakis, Distributed LMS for consensus-based in-network adaptive processing, *IEEE Transactions on Signal Processing*, vol. 57, no. 6, pp. 2365C2382, 2009.

[10] N. Takahashi, I. Yamada, and A. H. Sayed, Diffusion Least-Mean squares with adaptive combiners: formulation and performance analysis, *IEEE Transactions on Signal Processing*, vol. 58, no. 9, pp. 4795C4810, 2010.

[11] R. Arablouei, K. Doğançay, S. Werner, et al, Adaptive distributed estimation based on recursive least-squares and partial diffusion, *IEEE Transactions on Signal Processing*, vol. 62, no. 14, pp. 3510-3522, 2014.

[12] C. Chen, Z. X. Liu, L. Guo, Performance analysis of distributed adaptive filters, *Communications in Information and Systems*, vol. 15, no. 4, pp. 453-476, 2015.
[13] C. Chen, Z. X. Liu, L. Guo, Performance bounds of distributed adaptive filters with cooperative correlated signals, *Science China Information Sciences*, vol. 59, no. 11, 112202, 2016.
[14] L. Guo, Time-Varying Stochastic systems, *Jilin Scientific and Technological Press*, 1993.
[15] Q. Y. Liu, Z. D. Wang, X. He, et al, On Kalman-consensus filtering with random link failures over sensor networks, *IEEE Transactions on Automatic Control*, vol. 63, no. 8, pp. 2701-2708, 2018.
[16] H. F. Chen and L. Guo, Strong consistency of parameter estimates for discrete-time stochastic sytems, *J. Sys. Sci. & Math. Scis.*, vol. 5, no. 2, pp. 81-93, 1985.
[17] H. F. Chen and L. Guo, Strong consistency of recursive identification by no use of persistent excitation condition, *Acta Mathematicae Applicatae Sinica*, vol. 2, no. 2, pp. 133-145, 1985.
[18] R. P. Agaev and P. Y. Chebotarev, The matrix of maxium out forests of a digraph and its applications, *Autom. Remote Control*, vol. 61, No. 9, pp. 1424-1450, 2000.
[19] G. H. Hardy, J. E. Littlewood and G. polya, Inequalities, *Cambridge University Press*, 1934.