Automatica 151 (2023) 110958

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Brief paper

Distributed sparse identification for stochastic dynamic systems under cooperative non-persistent excitation condition ☆

Die Gan [a,b,c], Zhixin Liu [a,b,*]

[a] Key Laboratory of Systems and Control, Institute of Systems Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, PR China
[b] School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, PR China
[c] Zhongguancun Laboratory, Beijing 100094, PR China

ARTICLE INFO

Article history:
Received 27 February 2022
Received in revised form 11 November 2022
Accepted 28 January 2023
Available online 2 March 2023

Keywords:
Distributed sparse least squares
Stochastic dynamic system
L_1-regularization
Cooperative non-persistent excitation

ABSTRACT

This paper considers the distributed sparse identification problem over wireless sensor networks such that all sensors cooperatively estimate the unknown sparse parameter vector of stochastic dynamic systems by using the local information from neighbors. A distributed sparse least squares algorithm is proposed by minimizing a local information criterion formulated as a linear combination of accumulative local estimation error and L_1-regularization term. The upper bound of the estimation error of the proposed algorithm is presented. Furthermore, by designing a suitable adaptive weighting coefficient based on the local observation data, the set convergence of zero elements with a finite number of observations is obtained under a cooperative non-persistent excitation condition. It is shown that the proposed distributed algorithm can work well in a cooperative way even though none of the individual sensors can fulfill the estimation task. Our theoretical results are obtained without relying on the independency assumptions of regression signals that have been commonly used in the existing literature. Thus, our results are expected to be applied to stochastic feedback systems. Finally, the numerical simulations are provided to demonstrate the effectiveness of our theoretical results.

1. Introduction

In recent years, wireless sensor networks (WSNs) have attracted increasing research attention because of their wide application in engineering systems including smart grids, biomedical health monitoring, target tracking and surveillance (Sayed et al., 2013). Distributed observation and data analysis are ubiquitous in WSNs, where sensors are interconnected to acquire and process the local information from neighbors to finish a common task. Due to various uncertainties in practical systems, the distributed identification problem over WSNs becomes one of the important topics where all the sensors collaboratively estimate an unknown parameter vector of interest by using local noisy measurements. Unlike the centralized method with a fusion center, the distributed scheme has the advantages of flexibility, robustness to node or link failures as well as reducing communication load and calculation pressure. Consequently, the theoretical analysis of distributed estimation or filtering algorithms based on several typical distributed strategies such as the incremental, the diffusion and the consensus strategies have been provided (Abdolee & Champagne, 2016; Battilotti et al., 2020; Liu et al., 2020).

In practical scenarios, there exist a large number of sparse systems (Bazerque & Giannakis, 2010; Vinga, 2021) where many elements in the parameter vector do not contribute or contribute marginally to the systems (i.e., these elements are zero or near-zero). How to infer the zero elements and identify the nonzero elements in the unknown parameter vector is an important issue in the investigation of sparse systems. Considerable progress has been made on the identification of zero and nonzero elements in an unknown sparse parameter vector (Chiuso & Pillonetto, 2014; Eksioglu, 2013; Zhao & Yu, 2006), which allows us to obtain a more reliable prediction model. One direction for the estimation of sparse signals is based on the compressed sensing (CS) theory (Baraniuk, 2007; Candès & Tao, 2005), and some estimation algorithms using CS are proposed (cf., Gan and Liu (2022b), Xie and Guo (2020) and Xu et al. (2015)) in which *a priori* knowledge about the sparsity of the unknown parameter

☆ This work was supported by the National Key R&D Program of China under Grant 2018YFA0703800, Natural Science Foundation of China under Grant T2293770, China Postdoctoral Science Foundation, China (2022M722926) and National Science Foundation of Shandong Province (ZR2020ZD26). The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Er-Wei Bai under the direction of Editor Alessandro Chiuso.

* Corresponding author at: Key Laboratory of Systems and Control, Institute of Systems Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, PR China.

E-mail addresses: gandie@amss.ac.cn (D. Gan), lzx@amss.ac.cn (Z. Liu).

https://doi.org/10.1016/j.automatica.2023.110958
0005-1098/

and the regression vectors are required. Another direction is the sparse optimization based on the regularization framework where the objective function is formulated as a combination of the prediction error with a penalty term. The well-known LASSO (the least absolute shrinkage and selection operator) is one of the classical algorithms to obtain the sparse signals (Tibshirani, 1996), and its variants and adaptive LASSO (Zou, 2006) are also studied. For the stochastic dynamic systems with a single sensor, the adaptive sparse estimation or filtering algorithms are studied by combining the recursive least squares (LS) and least mean squares (LMS) with regularization term (Chen et al., 2009; Zhao et al., 2020).

With the development of sensor networks, some distributed adaptive sparse estimation algorithms have been proposed, and the corresponding stability and convergence analysis are also investigated under some signal conditions. For example, Di Lorenzo and Sayed (2013) provided the convergence and mean-square performance analysis for the distributed LMS algorithm regularized by convex penalties where the assumption of independent regressors is required. Huang and Li (2015) presented theoretical analysis on the mean and mean-square performance of the distributed sparse total LS algorithm under the condition that the input signals are independent and identically distributed (i.i.d.). Shiri et al. (2018) analyzed the mean stability of distributed quasi-sparse affine projection algorithm with independent regression vectors. Huang et al. (2020) analyzed the mean stability of the sparse diffusion LMS algorithm for two regularization terms with independent regression vectors. However, for the typical models such as ARMAX (autoregressive moving-average with exogenous input) model and Hammerstein system, the regressors are often generated by the past input and output signals, so it is hard for them to satisfy the aforementioned independency assumptions.

In order to relax the independency assumption of the regressors, some attempts are made for the distributed adaptive estimation or filtering algorithms. For the unknown time-invariant parameter vector, Gan and Liu (2019) proposed a distributed stochastic gradient algorithm, and established the strong consistency of the proposed algorithm under a cooperative excitation condition. Xie et al. (2021) studied the convergence of the diffusion LS algorithm. For the time-varying parameter vector, Xie and Guo (2018) provided a cooperative information condition to guarantee the stability of the consensus-based LMS adaptive filters. Moreover, Gan et al. (2021) introduced the collective random observability condition and provided the stability analysis of the distributed Kalman filter algorithm. Nevertheless, these asymptotical results are established as the number of the observation data obtained by sensors tends to infinity, which may not be suitable for the sparse identification problem with limited observation data. Kaiser et al. (2018) combined sparse identification of nonlinear dynamics with model predictive control in limited data for the single sensor case. The effectiveness of the proposed method was verified by some experiments without rigorous theoretical analysis.

Inspired by Zhao et al. (2020) where a sparse identification algorithm for a single sensor case is put forward to infer the set of zero elements with finite observations, we develop a distributed adaptive sparse LS algorithm over sensor networks such that all sensors can cooperatively identify the unknown parameter vector and infer the zero elements with a finite number of observations. The main contributions can be summarized as follows:

- We first introduce a local information criterion for each sensor which is formulated as a linear combination of local estimation errors with L_1-regularization term. By minimizing this criterion, a distributed adaptive sparse identification algorithm is proposed. The upper bound of the estimation error is established, which can be degenerated to the results of the classical distributed LS algorithm (Xie et al., 2021) when the weighting coefficients are equal to zero.
- Then, we introduce a cooperative non-persistent excitation condition on the regressors, under which the distributed sparse LS algorithm can cooperatively identify the set of zero elements with finite observations by properly choosing the weighting coefficients. We remark that the key difference between the proposed algorithm and those in distributed sparse optimization framework (e.g., Di Lorenzo and Sayed (2013)) lies in that the weighting coefficients are generated from the local observation sequences. The cooperative excitation condition is much weaker than the widely used persistent excitations (cf., Chen et al. (2015, 2014) and Zhang et al. (2021)) and the regularity condition (Zou, 2006).
- Different from most existing results on the distributed sparse algorithms, our theoretical results are obtained without relying on the independency assumptions of regression signals by virtue of powerful techniques including martingale theory, stochastic Lyapunov functions and convex optimization methods, which makes it possible for applications to the stochastic feedback systems. We also reveal that the whole sensor network can cooperatively accomplish the estimation task, even if any individual sensor cannot due to lack of necessary information (Zhao et al., 2020).

The remainder of this paper is organized as follows. In Section 2, we give the problem formulation of this paper; Section 3 presents the main results of the paper; the proofs of the main results are given in Section 4. A simulation example is provided in Section 5. Finally, we conclude the paper with some remarks in Section 6.

2. Problem formulation

2.1. Basic notations

In this paper, for an m-dimensional vector $\boldsymbol{x}$, its L_p-norm is defined as $\|\boldsymbol{x}\|_p = (\sum_{j=1}^m |\boldsymbol{x}(j)|^p)^{1/p}$ $(1 \le p < \infty)$, where $\boldsymbol{x}(j)$ denotes the jth element of $\boldsymbol{x}$. For $p = 1$, $\|\boldsymbol{x}\|_1$ is the sum of absolute values of all the elements in $\boldsymbol{x}$; and for $p = 2$, $\|\boldsymbol{x}\|_2$ is the Euclidean norm, we simply write $\|\cdot\|_2$ as $\|\cdot\|$. For an $m \times m$-dimensional real matrix $\boldsymbol{A}$, we use $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ to denote the largest and smallest eigenvalues of the matrix. $\|\boldsymbol{A}\|$ denotes the Euclidean norm, i.e., $\|\boldsymbol{A}\| = (\lambda_{\max}(\boldsymbol{A}\boldsymbol{A}^T))^{\frac{1}{2}}$ where the notation T denotes the transpose operator; $\|\boldsymbol{A}\|_F$ denotes the Frobenius norm, i.e., $\|\boldsymbol{A}\|_F = (tr(\boldsymbol{A}^T\boldsymbol{A}))^{\frac{1}{2}}$, where the notation $tr(\cdot)$ denotes the trace of the corresponding matrix. For a symmetric matrix $\boldsymbol{A}$, if all eigenvalues of $\boldsymbol{A}$ are positive (or nonnegative), then it is a positive (semipositive) definite matrix, and we denote it as $\boldsymbol{A} > 0$ (≥ 0). If all elements of a matrix $\boldsymbol{A} = \{a_{ij}\} \in \mathbb{R}^{n\times n}$ are nonnegative, then it is a nonnegative matrix, and furthermore if $\sum_{j=1}^n a_{ij} = 1$ holds for all $i \in \{1, \dots, n\}$, then it is called a stochastic matrix. For any two positive scalar sequences $\{a_k\}$ and $\{b_k\}$, by $a_k = O(b_k)$ we mean that there exists a constant $C > 0$ independent of k such that $a_k \le Cb_k$ holds for all $k \ge 0$, and by $a_k = o(b_k)$ we mean that $\lim_{k\to\infty} a_k/b_k = 0$.

2.2. Graph theory

We consider a sensor network with n sensors. The communication between sensors are usually modeled as an undirected weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$, where $\mathcal{V} = \{1, 2, 3, \dots, n\}$ is the set of sensors (or nodes), $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the edge set, and $\mathcal{A} = \{a_{ij}\} \in \mathbb{R}^{n\times n}$ is the weighted adjacency matrix. The elements of the adjacency matrix $\mathcal{A}$ satisfy $a_{ij} > 0$ if $(i, j) \in \mathcal{E}$ and $a_{ij} = 0$ otherwise. Here we assume that the matrix $\mathcal{A}$ is a symmetric and stochastic matrix. For the sensor i, the set of its neighbors is denoted as $N_i = \{j \in \mathcal{V} | (i, j) \in \mathcal{E}\}$, and the sensor i belongs to N_i.

The sensor i can communicate information with its neighboring sensors. A path of length ℓ is a sequence of nodes $\{i_1,\ldots,i_\ell,i_{\ell+1}\}$ such that $(i_h,i_{h+1})\in\mathcal{E}$ with $1\le h\le\ell$. The graph $\mathcal{G}$ is called connected if there is a path between any two sensors. The diameter $D_{\mathcal{G}}$ of the graph $\mathcal{G}$ is defined as the maximum shortest path length between any two sensors.

2.3. Observation model

In this paper, we consider the parameter identification problem in a network consisting of n sensors labeled $1,\ldots,n$. Assume that the data $\{y_{t,i},\boldsymbol{\varphi}_{t,i},t=1,2,\ldots\}$ collected by the sensor i obeys the following discrete-time stochastic regression model,

$$y_{t+1,i}=\boldsymbol{\varphi}_{t,i}^T\boldsymbol{\theta}+w_{t+1,i},\ \ t=0,1,2,\ldots, \tag{1}$$

where $y_{t,i}$ is the scalar observation or output of the sensor i at time t, $\boldsymbol{\varphi}_{t,i}$ is the m-dimensional stochastic regression vector, $\boldsymbol{\theta}\in\mathbb{R}^m$ is an unknown m-dimensional parameter to be estimated, and $\{w_{t,i}\}$ is the noise sequence. The above model (1) includes many parameterized systems, such as ARX system and Hammerstein system. We further denote the parameter vector $\boldsymbol{\theta}$ and the index set of its zero elements by

$$\begin{aligned}&\boldsymbol{\theta}\triangleq(\boldsymbol{\theta}(1),\ldots,\boldsymbol{\theta}(m))^T,\\&H^*\triangleq\{l\in\{1,\ldots,m\}|\boldsymbol{\theta}(l)=0\}.\end{aligned} \tag{2}$$

Our problem is to design a distributed adaptive estimation algorithm such that all sensors cooperatively infer the set H^* in a finite number of steps and identify the unknown parameter $\boldsymbol{\theta}$ by using stochastic regression vectors and the observation signals from its neighbors, i.e., $\{\boldsymbol{\varphi}_{k,j},y_{k+1,j}\}_{k=1}^t$ $(j\in N_i)$.

3. The main results

3.1. Parameter convergence

Before designing the algorithm to cooperatively estimate the unknown parameter vector and infer the set H^*, we first introduce the following classical distributed least squares algorithm to estimate the unknown parameter $\boldsymbol{\theta}$ in (2), i.e.,

$$\boldsymbol{\theta}_{t+1,i}=\boldsymbol{P}_{t+1,i}\left(\sum_{j=1}^n\sum_{k=0}^t a_{ij}^{(t+1-k)}\boldsymbol{\varphi}_{k,j}y_{k+1,j}\right), \tag{3}$$

where $\boldsymbol{P}_{t+1,i}=\left(\sum_{j=1}^n\sum_{k=0}^t a_{ij}^{(t+1-k)}\boldsymbol{\varphi}_{k,j}\boldsymbol{\varphi}_{k,j}^T\right)^{-1}$ and $a_{ij}^{(t+1-k)}$ is the ith row, jth column entry of the matrix $\mathcal{A}^{t+1-k}$ with $\mathcal{A}$ being the weighted adjacency matrix. It is clear that the matrix $\boldsymbol{P}_{t+1,i}$ can be equivalently written as the following recursive form,

$$\boldsymbol{P}_{t+1,i}^{-1}=\sum_{j\in N_i}a_{ij}(\boldsymbol{P}_{t,j}^{-1}+\boldsymbol{\varphi}_{t,j}\boldsymbol{\varphi}_{t,j}^T). \tag{4}$$

Thus, the algorithm (3) can also have the following recursive expression,

$$\boldsymbol{\theta}_{t+1,i}=\boldsymbol{P}_{t+1,i}\sum_{j\in N_i}a_{ij}(\boldsymbol{P}_{t,j}^{-1}\boldsymbol{\theta}_{t,j}+\boldsymbol{\varphi}_{t,j}y_{t+1,j}). \tag{5}$$

Note that in the above derivation, we use the inverse of the matrix $\sum_{j=1}^n\sum_{k=0}^t a_{ij}^{(t+1-k)}\boldsymbol{\varphi}_{k,j}\boldsymbol{\varphi}_{k,j}^T$ which is usually not satisfied for small t. To solve this problem, we take the initial matrix $\boldsymbol{P}_{0,i}$ to be positive definite. By (4), we have

$$\boldsymbol{P}_{t+1,i}^{-1}=\sum_{j=1}^n\sum_{k=0}^t a_{ij}^{(t+1-k)}\boldsymbol{\varphi}_{k,j}\boldsymbol{\varphi}_{k,j}^T+\sum_{j=1}^n a_{ij}^{(t+1)}\boldsymbol{P}_{0,j}^{-1}. \tag{6}$$

This modification will not affect the analysis of the asymptotic properties of the estimate of the distributed least squares algorithm.

In fact, the algorithm (5) can be obtained by minimizing the following linear combination of the estimation error $\sigma_{t+1,i}(\boldsymbol{\beta})$ between the observation signals and the prediction of the local neighbors,

$$\sigma_{t+1,i}(\boldsymbol{\beta})=\sum_{j\in N_i}a_{ij}\left(\sigma_{t,j}(\boldsymbol{\beta})+[y_{t+1,j}-\boldsymbol{\beta}^T\boldsymbol{\varphi}_{t,j}]^2\right), \tag{7}$$

with $\sigma_{0,i}(\boldsymbol{\beta})=0$. That is, $\boldsymbol{\theta}_{t+1,i}\triangleq\arg\min_{\boldsymbol{\beta}}\sigma_{t+1,i}(\boldsymbol{\beta})$.

By a simple calculation, (7) is equivalent to

$$\sigma_{t+1,i}(\boldsymbol{\beta})=\sum_{j=1}^n\sum_{k=0}^t a_{ij}^{(t+1-k)}[y_{k+1,j}-\boldsymbol{\beta}^T\boldsymbol{\varphi}_{k,j}]^2. \tag{8}$$

It is shown by Xie et al. (2021) that the distributed least squares algorithm (5) can generate a consistent estimate for the unknown parameter when the number of data tends to infinity. However, for the sparse unknown parameter vectors (i.e., there are many zero elements in $\boldsymbol{\theta}$), it is hard to infer the zero elements in a finite step due to the limitation of observations in practice. In order to solve this issue, we introduce the following local information criterion with L_1-regularization to identify the unknown sparse parameters and infer the set H^*,

$$J_{t+1,i}(\boldsymbol{\beta})=\sigma_{t+1,i}(\boldsymbol{\beta})+\alpha_{t+1,i}\|\boldsymbol{\beta}\|_1, \tag{9}$$

where $\|\cdot\|_1$ is the L_1-norm, $\alpha_{t+1,i}$ is the weighting coefficient chosen to satisfy $\alpha_{t+1,i}=o(\lambda_{\min}(\boldsymbol{P}_{t+1,i}^{-1}))$, and $\sigma_{t+1,i}(\boldsymbol{\beta})$ is recursively defined by (7). For the sensor i, we can obtain the following distributed sparse LS algorithm to estimate the unknown parameter $\boldsymbol{\theta}$ by minimizing $J_{t+1,i}(\boldsymbol{\beta})$, i.e.,

$$\boldsymbol{\beta}_{t+1,i}=\arg\min_{\boldsymbol{\beta}}J_{t+1,i}(\boldsymbol{\beta}). \tag{10}$$

Remark 1. We know that the L_0-regularization methods are ideal for variable selection in the sense of yielding the most sparse variables, but the computational complexity for solving L_0 minimization problem is NP-hard in general (cf., Candès and Tao (2005)). For L_2-regularization methods, the objective functions are smooth, but the solutions of L_2-regularization do not possess the sparse property. While the L_1-regularization can generate the sparse solutions, and it leads to a convex optimization problem that is easy to be solved by using the typical algorithms such as basic pursuit and interior-point algorithms (see e.g., Gill et al. (2011) and Kim et al. (2007)).

Remark 2. For the sensor i, the coefficients $\alpha_{t+1,i}$ in (9) can be dynamically adjusted by using the local observation sequence $\{\boldsymbol{\varphi}_{k,j},y_{k+1,j},j\in N_i\}_{k=1}^t$, which makes (9) be the adaptive LASSO (cf., Zou (2006)). We show that by properly choosing the coefficient $\alpha_{t+1,i}$, we can identify the set of the zero elements in the unknown sparse parameter vector $\boldsymbol{\theta}$ with a finite number of observations (see Theorem 2).

In the following, we will first investigate the upper bound of the estimation error generated by (10), which provides the basis for the set convergence of zero elements. For this purpose, we need to introduce the following assumptions on the network topology and the observation noise.

Assumption 1. The communication graph $\mathcal{G}$ is connected.

Remark 3. For the weighted adjacency matrix $\mathcal{A}$ of the graph $\mathcal{G}$, we denote $\mathcal{A}^l\triangleq(a_{ij}^{(l)})$ with $l\ge1$. By the theory of product of stochastic matrices, we see that under Assumption 1, $\mathcal{A}^l$ is a positive matrix for $l\ge D_{\mathcal{G}}$, i.e., for any i and j, $a_{ij}^{(l)}>0$.

Assumption 2. For any $i \in \{1, \ldots, n\}$, the noise sequence $\{w_{k,i}, \mathscr{F}_k\}$ is a martingale difference, and there exists a constant $\delta > 2$ such that $\sup_{k\geq 0} E[|w_{k+1,i}|^{\delta}|\mathscr{F}_k] < \infty$, a.s., where $\mathscr{F}_t = \sigma\{\boldsymbol{\varphi}_{k,i}, w_{k,i}, k \leq t, i = 1, \ldots, n\}$ is a sequence of nondecreasing σ-algebras and $E[\cdot|\cdot]$ denotes the conditional expectation operator.

We can verify that the i.i.d. zero-mean bounded or Gaussian noise $\{w_{k,i}\}$ which are independent of the regressors can satisfy Assumption 2.

Assume that there are d nonzero elements in the unknown parameter vector $\boldsymbol{\theta}$. Without loss of generality, we assume $\boldsymbol{\theta} = (\boldsymbol{\theta}(1), \ldots, \boldsymbol{\theta}(d), \boldsymbol{\theta}(d+1), \ldots, \boldsymbol{\theta}(m))^T$ with $\boldsymbol{\theta}(l) \neq 0, l = 1, \ldots, d$, and $\boldsymbol{\theta}(j) = 0, j = d+1, \ldots, m$. For the estimate $\boldsymbol{\beta}_{t+1,i}$ obtained by the distributed sparse LS algorithm (10), we denote the estimate error as

$$\widetilde{\boldsymbol{\beta}}_{t+1,i} = \boldsymbol{\beta}_{t+1,i} - \boldsymbol{\theta}. \tag{11}$$

Then we have the following result concerning the upper bound of the estimation error $\widetilde{\boldsymbol{\beta}}_{t,i}$.

Theorem 1. *Let $\boldsymbol{P}^{-1}_{t+1,i}$ be generated by (4) with arbitrarily initial matrix $\boldsymbol{P}_{0,i} > 0$. Then under Assumptions 1 and 2, we have for all $i \in \{1, \ldots, n\}$*

$$\|\widetilde{\boldsymbol{\beta}}_{t+1,i}\| = O\Big(\frac{\alpha_{t+1,i}}{\lambda_{\min}(\boldsymbol{P}^{-1}_{t+1,i})} + \sqrt{\frac{\log r_t}{\lambda_{\min}(\boldsymbol{P}^{-1}_{t+1,i})}}\Big), \text{ a.s.} \tag{12}$$

where $r_t = \sum_{i=1}^{n}\sum_{k=0}^{t}\|\boldsymbol{\varphi}_{k,i}\|^2$.

The proof of Theorem 1 is provided in Section 4.1.

Remark 4. By (6), we have for $t \geq D_{\mathcal{G}}$,

$$\lambda_{\min}(\boldsymbol{P}^{-1}_{t+1,i}) \geq a_{\min}\lambda^{n,t}_{\min}, \tag{13}$$

where $a_{\min} \triangleq \min_{i,j\in\mathcal{V}} a^{(D_{\mathcal{G}})}_{ij} > 0$ and $\lambda^{n,t}_{\min} = \lambda_{\min}\Big\{\sum_{j=1}^{n}\boldsymbol{P}^{-1}_{0,j} + \sum_{j=1}^{n}\sum_{k=0}^{t-D_{\mathcal{G}}+1}\boldsymbol{\varphi}_{k,j}\boldsymbol{\varphi}^T_{k,j}\Big\}$. From Theorem 1, if the coefficient $\alpha_{t+1,i}$ is chosen to satisfy $\alpha_{t+1,i} = o(\lambda_{\min}(\boldsymbol{P}^{-1}_{t+1,i}))$ and the regression vectors satisfy the weakest possible cooperative excitation condition $\log r_t = o(\lambda^{n,t}_{\min})$ (cf., Xie et al. (2021)), then the almost sure convergence of the distributed sparse LS algorithm can be obtained, i.e., $\boldsymbol{\beta}_{t+1,i} \xrightarrow[t\to\infty]{} \boldsymbol{\theta}$. Theorem 1 can be degenerated to the results of the classical distributed LS algorithm in Xie et al. (2021) when $\alpha_{t+1,i}$ is equal to zero.

3.2. Set convergence

In last subsection, we have obtained the asymptotic results concerning the parameter convergence. In the following, we introduce the cooperative non-persistent excitation condition to study the convergence of the sets of zero elements in the unknown sparse parameter vector with a finite number of observations, which is different from the asymptotic analysis given in last subsection.

Assumption 3 (*Cooperative Non-persistent Excitation Condition*)**.** The following condition is satisfied,

$$\frac{r_t}{\lambda^{n,t}_{\min}}\sqrt{\frac{\log(r_t)}{\lambda^{n,t}_{\min}}} \xrightarrow[t\to\infty]{} 0, \quad \text{a.s.} \tag{14}$$

where r_t and $\lambda^{n,t}_{\min}$ are respectively defined in Theorem 1 and Remark 4.

Remark 5. For the single sensor case with $n = 1$ and $D_{\mathcal{G}} = 1$, the condition (14) reduces to the excitation condition given by Zhao et al. (2020). Assumption 3 reveals the cooperative effect of multiple sensors in the sense that the condition (14) can make it possible for Algorithm 1 to estimate the unknown parameter $\boldsymbol{\theta}$ and the sets of zero elements by the cooperation of multiple sensors even if any individual sensor cannot due to lack of adequate excitation, which is also shown in the simulation example given in Section 5.

Inspired by Zhao et al. (2020), we propose the following distributed sparse adaptive algorithm (Algorithm 1) to identify the set of zero elements with a finite number of observations by choosing $\alpha_{t,i}$ adaptively.

Algorithm 1

Step 1: Based on $\{\boldsymbol{\varphi}_{k,j}, y_{k+1,j}\}_{k=0}^{t}$ $(j \in N_i)$, begin with an initial vector $\boldsymbol{\theta}_{0,i}$ and an initial matrix $\boldsymbol{P}_{0,i} > 0$, $(i = 1, 2\ldots, n)$, compute the matrix $\boldsymbol{P}^{-1}_{t+1,i}$ defined by (4) and the local estimate $\boldsymbol{\theta}_{t+1,i}$ of $\boldsymbol{\theta}$ by (5), and further compute $\hat{\boldsymbol{\theta}}_{t+1,i}(l)$ according to

$$\begin{aligned}&\hat{\boldsymbol{\theta}}_{t+1,i}(l)\\ &= \boldsymbol{\theta}_{t+1,i}(l) + \mathrm{sgn}(\boldsymbol{\theta}_{t+1,i}(l))\sqrt{\frac{\log(\lambda_{\max}(\boldsymbol{P}^{-1}_{t+1,i}))}{\lambda_{\min}(\boldsymbol{P}^{-1}_{t+1,i})}},\end{aligned} \tag{15}$$

Step 2: Choose a positive sequence $\{\alpha_{k,i}\}_{k=1}^{t+1}$ satisfying

$$\begin{aligned}&\alpha_{k,i} = o(\lambda_{\min}(\boldsymbol{P}^{-1}_{k,i})),\\ &\lambda_{\max}(\boldsymbol{P}^{-1}_{k,i})\sqrt{\frac{\log(\lambda_{\max}(\boldsymbol{P}^{-1}_{k,i}))}{\lambda_{\min}(\boldsymbol{P}^{-1}_{k,i})}} = o(\alpha_{k,i}).\end{aligned} \tag{16}$$

Step 3: Optimize the convex local objective function,

$$\bar{J}_{t+1,i}(\boldsymbol{\xi}) = \sigma_{t+1,i}(\boldsymbol{\xi}) + \alpha_{t+1,i}\sum_{l=1}^{m}\frac{1}{|\hat{\boldsymbol{\theta}}_{t+1,i}(l)|}|\boldsymbol{\xi}(l)| \tag{17}$$

with $\sigma_{t+1,i}(\boldsymbol{\xi})$ defined in (7), and obtain

$$\begin{aligned}\boldsymbol{\xi}_{t+1,i} &= (\boldsymbol{\xi}_{t+1,i}(1), \cdots, \boldsymbol{\xi}_{t+1,i}(m))^T\\ &\triangleq \arg\min_{\boldsymbol{\xi}} \bar{J}_{t+1,i}(\boldsymbol{\xi}),\end{aligned} \tag{18}$$

$$H_{t+1,i} \triangleq \{l = 1, \cdots, m|\boldsymbol{\xi}_{t+1,i}(l) = 0\}. \tag{19}$$

In fact, $\alpha_{k,i}$ in Step 2 can be taken as

$$\sqrt{\lambda_{\max}(\boldsymbol{P}^{-1}_{k,i})\lambda_{\min}(\boldsymbol{P}^{-1}_{k,i})\sqrt{\frac{\log(\lambda_{\max}(\boldsymbol{P}^{-1}_{k,i}))}{\lambda_{\min}(\boldsymbol{P}^{-1}_{k,i})}}}$$

since by Assumption 3, (6) and (12) we have

$$\frac{\lambda_{\max}(\boldsymbol{P}^{-1}_{k,i})}{\lambda_{\min}(\boldsymbol{P}^{-1}_{k,i})}\sqrt{\frac{\log(\lambda_{\max}(\boldsymbol{P}^{-1}_{k,i}))}{\lambda_{\min}(\boldsymbol{P}^{-1}_{k,i})}} = o(1).$$

In the convex objective function (17), different components in $\boldsymbol{\xi}$ are assigned different weights, which is an adaptive LASSO estimator since the weights $\alpha_{t+1,i}/\hat{\boldsymbol{\theta}}_{t+1,i}(l)$ are generated from the local observation sequence $\{\boldsymbol{\varphi}_{k,j}, y_{k+1,j}, j \in N_i\}_{k=1}^{t}$. The $\hat{\boldsymbol{\theta}}_{t+1,i}(l)$ appearing in the denominator satisfies that $|\hat{\boldsymbol{\theta}}_{t+1,i}(l)| \geq \sqrt{\frac{\log(\lambda_{\max}(\boldsymbol{P}^{-1}_{t+1,i}))}{\lambda_{\min}(\boldsymbol{P}^{-1}_{t+1,i})}} > 0$, which makes (17) well defined. Moreover, if $\hat{\boldsymbol{\theta}}_{t+1,i}(l) \to 0$ for some $l = 1, \ldots, m$ and hence $1/\hat{\boldsymbol{\theta}}_{t+1,i}(l) \to \infty$, then the corresponding minimizer $\xi_{t+1,i}(l)$ should be exactly zero. This provides an intuitive explanation for the sparse solution of Algorithm 1 with a finite number of observations. The set $H_{t+1,i}$ generated from the convex optimization problem (18) serves as the estimate for the set H^* defined in (2).

For the set $H_{t,i}$ obtained by (19), we get the following finite time convergence result, which shows that the set of zero elements in $\boldsymbol{\theta}$ can be correctly identified with a finite number of observations.

Theorem 2 (*Set Convergence*). *Under Assumptions 1–3, if $\log r_t = O(\log r_{t-D_{\mathcal{G}}+1})$, then there exists a positive integer T_0 (which may depend on the sample ω) such that for all $i \in \{1,\dots,n\}$*

$$\xi_{t+1,i}(d+1) = \cdots = \xi_{t+1,i}(m) = 0, \quad t \geq T_0.$$

That is, $H_{t+1,i} = H^*$ for $t \geq T_0$, where H^* and $H_{t+1,i}$ are defined in (2) and (19).

The detailed proof of Theorem 2 is given in Section 4.2.

Remark 6. The condition $\log(r_t) = O(\log(r_{t-D_{\mathcal{G}}+1}))$ of Theorem 2 means that the growth rate of regression vector is not explosive. For some typical cases of regression vectors $\{\boldsymbol{\varphi}_k^i\}$ such as the bounded sequence or even sequences with exponential growth rate, and the i.i.d. sequence, the condition $\log(r_t) = O(\log(r_{t-D_{\mathcal{G}}+1}))$ can be easily verified.

Remark 7. From Theorem 2 (also Theorem 1), we see that the parameter convergence and set convergence results in this paper are derived without using the independency assumption on the regression vectors, which makes it possible to apply our algorithm to practical feedback systems.

4. Proofs of the main results

In order to prove the main theorems of the paper, we first give the following preliminary lemma (Gan & Liu, 2022a; Xie et al., 2021).

Lemma 1. *Under Assumptions 1 and 2, we have the following results for all i,*

(1) $$\sum_{i=1}^{n} \|\widetilde{\boldsymbol{\theta}}_{t,i}\|^2 = O\left(\frac{\log r_t}{\lambda_{\min}^{n,t}}\right),$$

(2) $$\left\| \boldsymbol{P}_{t,i}^{\frac{1}{2}} \left(\sum_{j=1}^{n}\sum_{k=0}^{t} a_{ij}^{(t+1-k)} \boldsymbol{\varphi}_{k,j} w_{k+1,j} \right) \right\| = O(\sqrt{\log(r_t)}).$$

where $\widetilde{\boldsymbol{\theta}}_{t+1,i} \triangleq \boldsymbol{\theta}_{t+1,i} - \boldsymbol{\theta}$ is the estimation error of the classical distributed LS algorithm (5).

4.1. Proof of Theorem 1

Proof. By noting that $\boldsymbol{\beta}_{t+1,i}$ is the minimizer of $J_{t+1,i}(\boldsymbol{\beta})$, it follows that

$$\begin{aligned} 0 &\geq J_{t+1,i}(\boldsymbol{\beta}_{t+1,i}) - J_{t+1,i}(\boldsymbol{\theta}) \\ &= J_{t+1,i}(\widetilde{\boldsymbol{\beta}}_{t+1,i} + \boldsymbol{\theta}) - J_{t+1,i}(\boldsymbol{\theta}). \end{aligned} \tag{20}$$

Since $\boldsymbol{\theta}(j) = 0$, $j = d+1,\dots,m$, by (1), (8) and (9), we have

$$\begin{aligned} &J_{t+1,i}(\widetilde{\boldsymbol{\beta}}_{t+1,i} + \boldsymbol{\theta}) = \sum_{j=1}^{n}\sum_{k=0}^{t} a_{ij}^{(t+1-k)} [w_{k+1,j} - \widetilde{\boldsymbol{\beta}}_{t+1,i}^T \boldsymbol{\varphi}_{k,j}]^2 \\ &+ \alpha_{t+1,i}\sum_{l=1}^{d} |\widetilde{\boldsymbol{\beta}}_{t+1,i}(l) + \boldsymbol{\theta}(l)| + \alpha_{t+1,i}\sum_{l=d+1}^{m} |\widetilde{\boldsymbol{\beta}}_{t+1,i}(l)| = \\ &\sum_{j=1}^{n}\sum_{k=0}^{t} a_{ij}^{(t+1-k)} w_{k+1,j}^2 + \widetilde{\boldsymbol{\beta}}_{t+1,i}^T \sum_{k=0}^{t} a_{ij}^{(t+1-k)} \boldsymbol{\varphi}_{k,j}\boldsymbol{\varphi}_{k,j}^T \widetilde{\boldsymbol{\beta}}_{t+1,i} \\ &- 2\widetilde{\boldsymbol{\beta}}_{t+1,i}^T \sum_{j=1}^{n}\sum_{k=0}^{t} a_{ij}^{(t+1-k)} \boldsymbol{\varphi}_{k,j} w_{k+1,j} \\ &+ \alpha_{t+1,i}\sum_{l=1}^{d} |\widetilde{\boldsymbol{\beta}}_{t+1,i}(l) + \boldsymbol{\theta}(l)| + \alpha_{t+1,i}\sum_{l=d+1}^{m} |\widetilde{\boldsymbol{\beta}}_{t+1,i}(l)|. \end{aligned} \tag{21}$$

Similarly, we have

$$\begin{aligned} &J_{t+1,i}(\boldsymbol{\theta}) \\ &= \sum_{j=1}^{n}\sum_{k=0}^{t} a_{ij}^{(t+1-k)} [y_{k+1,j} - \boldsymbol{\theta}^T \boldsymbol{\varphi}_{k,j}]^2 + \alpha_{t+1,i}\sum_{l=1}^{d} |\boldsymbol{\theta}(l)| \\ &= \sum_{j=1}^{n}\sum_{k=0}^{t} a_{ij}^{(t+1-k)} w_{k+1,j}^2 + \alpha_{t+1,i}\sum_{l=1}^{d} |\boldsymbol{\theta}(l)|. \end{aligned} \tag{22}$$

Hence by (21) and (22), we have

$$\begin{aligned} &J_{t+1,i}(\widetilde{\boldsymbol{\beta}}_{t+1,i} + \boldsymbol{\theta}) - J_{t+1,i}(\boldsymbol{\theta}) \\ &\geq \widetilde{\boldsymbol{\beta}}_{t+1,i}^T \sum_{k=0}^{t} a_{ij}^{(t+1-k)} \boldsymbol{\varphi}_{k,j}\boldsymbol{\varphi}_{k,j}^T \widetilde{\boldsymbol{\beta}}_{t+1,i} \\ &\quad - 2\widetilde{\boldsymbol{\beta}}_{t+1,i}^T \sum_{j=1}^{n}\sum_{k=0}^{t} a_{ij}^{(t+1-k)} \boldsymbol{\varphi}_{k,j} w_{k+1,j} \\ &\quad + \alpha_{t+1,i}\sum_{l=1}^{d} (|\widetilde{\boldsymbol{\beta}}_{t+1,i}(l) + \boldsymbol{\theta}(l)| - |\boldsymbol{\theta}(l)|) \\ &\triangleq M_{t+1,i}^{(1)} - 2M_{t+1,i}^{(2)} + M_{t+1,i}^{(3)}. \end{aligned} \tag{23}$$

In the following, we estimate $M_{t+1,i}^{(1)}$, $M_{t+1,i}^{(2)}$ and $M_{t+1,i}^{(3)}$ separately. Denote $\boldsymbol{V}_{t+1,i} = \boldsymbol{P}_{t+1,i}^{-\frac{1}{2}} \widetilde{\boldsymbol{\beta}}_{t+1,i}$. By Lemma 1, we have

$$\begin{aligned} |M_{t+1,i}^{(2)}| &= \left| \widetilde{\boldsymbol{\beta}}_{t+1,i}^T \boldsymbol{P}_{t+1,i}^{-\frac{1}{2}} \boldsymbol{P}_{t+1,i}^{\frac{1}{2}} \sum_{j=1}^{n}\sum_{k=0}^{t} a_{ij}^{(t+1-k)} \boldsymbol{\varphi}_{k,j} w_{k+1,j} \right| \\ &= O(\sqrt{\log(r_t)}) \|\boldsymbol{V}_{t+1,i}\|. \end{aligned}$$

Hence, there exists a positive constant c_1 such that for large t, $M_{t+1,i}^{(1)} - 2M_{t+1,i}^{(2)} \geq \frac{\|\boldsymbol{V}_{t+1,i}\|^2}{2} - c_1\sqrt{\log(r_t)}\|\boldsymbol{V}_{t+1,i}\|$. By C_r-inequality, we have $|M_{t+1,i}^{(3)}| \leq \alpha_{t+1,i}\sum_{l=1}^{d} |\widetilde{\boldsymbol{\beta}}_{t+1,i}(l)| \leq \alpha_{t+1,i}\sqrt{d}\|\widetilde{\boldsymbol{\beta}}_{t+1,i}\|$. Hence by (20) and (23), we have for large t, $0 \geq \frac{\|\boldsymbol{V}_{t+1,i}\|^2}{2} - c_1\sqrt{\log(r_t)}\|\boldsymbol{V}_{t+1,i}\| - \sqrt{d}\alpha_{t+1,i}\|\widetilde{\boldsymbol{\beta}}_{t+1,i}\|$, which implies that $\|\boldsymbol{V}_{t+1,i}\| \leq \sqrt{c_1^2 \log r_t + 2\sqrt{d}\alpha_{t+1,i}\|\widetilde{\boldsymbol{\beta}}_{t+1,i}\|} + \sqrt{c_1 \log r_t}$. Note that $\|\boldsymbol{V}_{t+1,i}\|^2 \geq \lambda_{\min}(\boldsymbol{P}_{t+1,i}^{-1})\|\widetilde{\boldsymbol{\beta}}_{t+1,i}\|^2$. Hence we obtain

$$\begin{aligned} &\left(\|\widetilde{\boldsymbol{\beta}}_{t+1,i}\| - \frac{2\sqrt{d}\alpha_{t+1,i}}{\lambda_{\min}(\boldsymbol{P}_{t+1,i}^{-1})} \right)^2 \\ &\leq \left(\frac{2\sqrt{d}\alpha_{t+1,i}}{\lambda_{\min}(\boldsymbol{P}_{t+1,i}^{-1})} \right)^2 + \frac{(2c_1^2 + 2c_1)\log r_t}{\lambda_{\min}(\boldsymbol{P}_{t+1,i}^{-1})}. \end{aligned} \tag{24}$$

Thus, we have $\|\widetilde{\boldsymbol{\beta}}_{t+1,i}\| = O\Big(\frac{\alpha_{t+1,i}}{\lambda_{\min}(\boldsymbol{P}_{t+1,i}^{-1})} + \sqrt{\frac{\log r_t}{\lambda_{\min}(\boldsymbol{P}_{t+1,i}^{-1})}}\Big)$, which completes the proof of the theorem. □

4.2. Proof of Theorem 2

Proof. Denote the estimation error between $\xi_{t+1,i}$ obtained by Algorithm 1 and $\boldsymbol{\theta}$ as $\widetilde{\xi}_{t+1,i} = \xi_{t+1,i} - \boldsymbol{\theta}$. By Assumption 3 and

Lemma 1, we see that the limits of $\boldsymbol{\theta}_{t+1,i}(l)$ and $\hat{\boldsymbol{\theta}}_{t+1,i}(l)$, $l = 1,\ldots,d$ are nonzero. Similar to the proof of Theorem 1, we can obtain $\|\widetilde{\boldsymbol{\xi}}_{t+1,i}\| = O\left(\frac{\alpha_{t+1,i}}{\lambda_{\min}(\boldsymbol{P}^{-1}_{t+1,i})} + \sqrt{\frac{\log r_t}{\lambda_{\min}(\boldsymbol{P}^{-1}_{t+1,i})}}\right)$. By the definition of $\widetilde{\boldsymbol{\xi}}_{t+1,i}$, it suffices to prove that there exists a positive integer T_0 such that for all $i \in \{1,\ldots,n\}$, $\widetilde{\boldsymbol{\xi}}_{t+1,i}(d+1) = \cdots = \widetilde{\boldsymbol{\xi}}_{t+1,i}(m) = 0$, $\quad t \geq T_0$. Otherwise, if for some $s_l \in \{d+1,\ldots,m\}$, some sensor i_0, and some subsequence $\{t_p\}_{p\geq 1}$ such that $\widetilde{\boldsymbol{\xi}}_{t_p+1,i_0}(s_l) \neq 0$, $p \geq 1$. Thus for $p \geq 1$, we have $\|\widetilde{\boldsymbol{\xi}}_{t_p+1,i_0}\| > 0$. Denote $\widetilde{\boldsymbol{\xi}}_{t_p+1,i_0} = (\widetilde{\boldsymbol{\xi}}^{(1)T}_{t_p+1,i_0}, \widetilde{\boldsymbol{\xi}}^{(2)T}_{t_p+1,i_0})^T$ and $\bar{\boldsymbol{\xi}}_{t_p+1,i_0} = (\widetilde{\boldsymbol{\xi}}^{(1)T}_{t_p+1,i_0}, \boldsymbol{0}^T)^T$, where $\widetilde{\boldsymbol{\xi}}^{(1)}_{t_p+1,i_0} \in \mathbb{R}^d$ and $\widetilde{\boldsymbol{\xi}}^{(2)}_{t_p+1,i_0} \in \mathbb{R}^{m-d}$. By noting that $\boldsymbol{\xi}_{t_p+1,i_0}$ is the minimizer of $\bar{J}_{t_p+1,i_0}(\boldsymbol{\xi})$ defined by (17), it follows that

$$\begin{aligned}0 &\geq \bar{J}_{t_p+1,i_0}(\boldsymbol{\xi}_{t_p+1,i_0}) - \bar{J}_{t_p+1,i_0}(\boldsymbol{\theta}+\bar{\boldsymbol{\xi}}_{t_p+1,i_0})\\ &= \bar{J}_{t_p+1,i_0}(\boldsymbol{\theta}+\widetilde{\boldsymbol{\xi}}_{t_p+1,i_0}) - \bar{J}_{t_p+1,i_0}(\boldsymbol{\theta}+\bar{\boldsymbol{\xi}}_{t_p+1,i_0}).\end{aligned} \tag{25}$$

Denote $\boldsymbol{\varphi}_{k,j} \triangleq (\boldsymbol{\varphi}^{(1)T}_{k,j}, \boldsymbol{\varphi}^{(2)T}_{k,j})^T$ and

$$\boldsymbol{\Psi}_{t+1,i} = \sum_{j=1}^{n}\sum_{k=0}^{t} a_{ij}^{(t+1-k)}\boldsymbol{\varphi}_{k,j}\boldsymbol{\varphi}^T_{k,j} \triangleq \begin{pmatrix}\boldsymbol{\Psi}^{(11)}_{t+1,i} & \boldsymbol{\Psi}^{(12)}_{t+1,i}\\ \boldsymbol{\Psi}^{(21)}_{t+1,i} & \boldsymbol{\Psi}^{(22)}_{t+1,i}\end{pmatrix}.$$

Similar to (21), we have for $\widetilde{\boldsymbol{\xi}}_{t_p+1,i_0}$

$$\begin{aligned}&\bar{J}_{t_p+1,i_0}(\boldsymbol{\theta}+\widetilde{\boldsymbol{\xi}}_{t_p+1,i_0}) - \sum_{j=1}^{n}\sum_{k=0}^{t_p} a_{i_0j}^{(t_p+1-k)} w^2_{k+1,j}\\ &= -2\widetilde{\boldsymbol{\xi}}^{(1)T}_{t_p+1,i_0}\sum_{j=1}^{n}\sum_{k=0}^{t_p} a_{i_0j}^{(t_p+1-k)}\boldsymbol{\varphi}^{(1)}_{k,j} w_{k+1,j}\\ &\quad - 2\widetilde{\boldsymbol{\xi}}^{(2)T}_{t_p+1,i_0}\sum_{j=1}^{n}\sum_{k=0}^{t_p} a_{i_0j}^{(t_p+1-k)}\boldsymbol{\varphi}^{(2)}_{k,j} w_{k+1,j}\\ &\quad + \widetilde{\boldsymbol{\xi}}^{(1)T}_{t_p+1,i_0}\boldsymbol{\Psi}^{(11)}_{t_p+1,i_0}\widetilde{\boldsymbol{\xi}}^{(1)}_{t_p+1,i_0} + \widetilde{\boldsymbol{\xi}}^{(2)T}_{t_p+1,i_0}\boldsymbol{\Psi}^{(21)}_{t_p+1,i_0}\widetilde{\boldsymbol{\xi}}^{(1)}_{t_p+1,i_0}\\ &\quad + \widetilde{\boldsymbol{\xi}}^{(1)T}_{t_p+1,i_0}\boldsymbol{\Psi}^{(12)}_{t_p+1,i_0}\widetilde{\boldsymbol{\xi}}^{(2)}_{t_p+1,i_0} + \widetilde{\boldsymbol{\xi}}^{(2)T}_{t_p+1,i_0}\boldsymbol{\Psi}^{(22)}_{t_p+1,i_0}\widetilde{\boldsymbol{\xi}}^{(2)}_{t_p+1,i_0}\\ &\quad + \alpha_{t_p+1,i_0}\sum_{l=1}^{d}\frac{1}{\hat{\boldsymbol{\theta}}_{t_p+1,i_0}(l)}|\widetilde{\boldsymbol{\xi}}_{t_p+1,i_0}(l) + \boldsymbol{\theta}(l)|\\ &\quad + \alpha_{t_p+1,i_0}\sum_{l=d+1}^{m}\frac{1}{|\hat{\boldsymbol{\theta}}_{t_p+1,i_0}(l)|}|\widetilde{\boldsymbol{\xi}}_{t_p+1,i_0}(l)|.\end{aligned} \tag{26}$$

By the definition of $\bar{\boldsymbol{\xi}}_{t_p+1,i_0}$, we have

$$\begin{aligned}&\bar{J}_{t_p+1,i_0}(\boldsymbol{\theta}+\bar{\boldsymbol{\xi}}_{t_p+1,i_0}) - \sum_{j=1}^{n}\sum_{k=0}^{t_p} a_{i_0j}^{(t_p+1-k)} w^2_{k+1,j}\\ &= -2\widetilde{\boldsymbol{\xi}}^{(1)T}_{t_p+1,i_0}\sum_{j=1}^{n}\sum_{k=0}^{t_p} a_{i_0j}^{(t_p+1-k)}\boldsymbol{\varphi}^{(1)}_{k,j} w_{k+1,j}\\ &\quad + \widetilde{\boldsymbol{\xi}}^{(1)T}_{t_p+1,i_0}\boldsymbol{\Psi}^{(11)}_{t_p+1,i_0}\widetilde{\boldsymbol{\xi}}^{(1)}_{t_p+1,i_0}\\ &\quad + \alpha_{t_p+1,i_0}\sum_{l=1}^{d}\frac{1}{|\hat{\boldsymbol{\theta}}_{t_p+1,i_0}(l)|}|\bar{\boldsymbol{\xi}}_{t_p+1,i_0}(l) + \boldsymbol{\theta}(l)|.\end{aligned} \tag{27}$$

By (26) and (27), we have

$$\begin{aligned}&\bar{J}_{t_p+1,i_0}(\boldsymbol{\theta}+\widetilde{\boldsymbol{\xi}}_{t_p+1,i_0}) - \bar{J}_{t_p+1,i_0}(\boldsymbol{\theta}+\bar{\boldsymbol{\xi}}_{t_p+1,i_0})\\ &= -2\widetilde{\boldsymbol{\xi}}^{(2)T}_{t_p+1,i_0}\sum_{j=1}^{n}\sum_{k=0}^{t_p} a_{i_0j}^{(t_p+1-k)}\boldsymbol{\varphi}^{(2)}_{k,j} w_{k+1,j}\\ &\quad + \widetilde{\boldsymbol{\xi}}^{(2)T}_{t_p+1,i_0}\boldsymbol{\Psi}^{(22)}_{t_p+1,i_0}\widetilde{\boldsymbol{\xi}}^{(2)}_{t_p+1,i_0} + \widetilde{\boldsymbol{\xi}}^{(1)T}_{t_p+1,i_0}\boldsymbol{\Psi}^{(12)}_{t_p+1,i_0}\widetilde{\boldsymbol{\xi}}^{(2)}_{t_p+1,i_0}\\ &\quad + \widetilde{\boldsymbol{\xi}}^{(2)T}_{t_p+1,i_0}\boldsymbol{\Psi}^{(21)}_{t_p+1,i_0}\widetilde{\boldsymbol{\xi}}^{(1)}_{t_p+1,i_0} + \sum_{l=d+1}^{m}\frac{\alpha_{t_p+1,i_0}}{|\hat{\boldsymbol{\theta}}_{t_p+1,i_0}(l)|}|\widetilde{\boldsymbol{\xi}}_{t_p+1,i_0}(l)|\\ &\triangleq -2I^{(1)}_{t_p+1,i_0} + I^{(2)}_{t_p+1,i_0} + I^{(3)}_{t_p+1,i_0} + I^{(4)}_{t_p+1,i_0} + I^{(5)}_{t_p+1,i_0}.\end{aligned} \tag{28}$$

In the following, we estimate $I^{(1)}_{t_p+1,i_0}$, $I^{(2)}_{t_p+1,i_0}$, $I^{(3)}_{t_p+1,i_0}$, $I^{(4)}_{t_p+1,i_0}$, $I^{(5)}_{t_p+1,i_0}$ separately. By (6), we have $\boldsymbol{P}^{-1}_{t+1,i} = \boldsymbol{\Psi}_{t+1,i} + \sum_{j=1}^{n} a_{ij}^{(t+1)}\boldsymbol{P}^{-1}_{0,j} \triangleq \begin{pmatrix}\boldsymbol{Q}^{(11)}_{t+1,i} & \boldsymbol{Q}^{(12)}_{t+1,i}\\ \boldsymbol{Q}^{(21)}_{t+1,i} & \boldsymbol{Q}^{(22)}_{t+1,i}\end{pmatrix}$. By Lemma 1, we have

$$\begin{aligned}|I^{(1)}_{t_p+1,i_0}| &= \Bigg|\widetilde{\boldsymbol{\xi}}^{(2)T}_{t_p+1,i_0}(\boldsymbol{Q}^{(22)}_{t_p+1,i_0})^{\frac{1}{2}}(\boldsymbol{Q}^{(22)}_{t_p+1,i_0})^{-\frac{1}{2}}\\ &\qquad \sum_{j=1}^{n}\sum_{k=0}^{t_p} a_{i_0j}^{(t_p+1-k)}\boldsymbol{\varphi}^{(2)}_{k,j} w_{k+1,j}\Bigg|\\ &= \|(\boldsymbol{Q}^{(22)}_{t_p+1,i_0})\|^{\frac{1}{2}}\|\widetilde{\boldsymbol{\xi}}^{(2)}_{t_p+1,i_0}\| O\left(\sqrt{\log r^{(2)}_{t_p}}\right),\end{aligned} \tag{29}$$

where $r^{(2)}_t \triangleq \sum_{i=1}^{n}\sum_{k=0}^{t}\|\boldsymbol{\varphi}^{(2)}_{k,i}\|^2$. Note that $\lambda_{\max}(\boldsymbol{Q}^{(22)}_{t_p+1,i_0}) \leq \lambda_{\max}(\boldsymbol{P}^{-1}_{t_p+1,i_0})$ and $\lambda_{\min}(\boldsymbol{Q}^{(22)}_{t_p+1,i_0}) \geq \lambda_{\min}(\boldsymbol{P}^{-1}_{t_p+1,i_0})$. Hence, we have $r^{(2)}_{t_p} \leq r_{t_p}$. We obtain that for large p and some positive constant c_2

$$\begin{aligned}&-2I^{(1)}_{t_p+1,i_0} + I^{(2)}_{t_p+1,i_0} \geq \lambda_{\min}(\boldsymbol{\Psi}^{(22)}_{t_p+1,i_0})\|\widetilde{\boldsymbol{\xi}}^{(2)}_{t_p+1,i_0}\|^2\\ &-c_2\|(\boldsymbol{Q}^{(22)}_{t_p+1,i_0})\|^{\frac{1}{2}}\|\widetilde{\boldsymbol{\xi}}^{(2)}_{t_p+1,i_0}\|\sqrt{\log r^{(2)}_{t_p}}\\ &\geq \frac{1}{2}\lambda_{\min}(\boldsymbol{Q}^{(22)}_{t_p+1,i_0})\|\widetilde{\boldsymbol{\xi}}^{(2)}_{t_p+1,i_0}\|^2\\ &\quad - c_2\|(\boldsymbol{Q}^{(22)}_{t_p+1,i_0})\|^{\frac{1}{2}}\|\widetilde{\boldsymbol{\xi}}^{(2)}_{t_p+1,i_0}\|\sqrt{\log r^{(2)}_{t_p}}\\ &\geq \frac{1}{2}\lambda_{\min}(\boldsymbol{P}^{-1}_{t_p+1,i_0})\|\widetilde{\boldsymbol{\xi}}^{(2)}_{t_p+1,i_0}\|^2\\ &\quad - c_2\sqrt{\lambda_{\max}(\boldsymbol{P}^{-1}_{t_p+1,i_0})}\|\widetilde{\boldsymbol{\xi}}^{(2)}_{t_p+1,i_0}\|\sqrt{\log r_{t_p}}.\end{aligned} \tag{30}$$

By Lemma 1, and based on the equivalence of norms in a finite dimensional space, we have

$$\begin{aligned}|I^{(3)}_{t_p+1,i_0}| &\leq c_3\|\widetilde{\boldsymbol{\xi}}^{(1)}_{t_p+1,i_0}\|\|\widetilde{\boldsymbol{\xi}}^{(2)}_{t_p+1,i_0}\|\|\boldsymbol{\Psi}^{(12)}_{t_p+1,i_0}\|_F\\ &\leq c_4\|\widetilde{\boldsymbol{\xi}}^{(1)}_{t_p+1,i_0}\|\|\widetilde{\boldsymbol{\xi}}^{(2)}_{t_p+1,i_0}\|\|\boldsymbol{\Psi}_{t_p+1,i_0}\|\\ &\leq c_4\|\widetilde{\boldsymbol{\xi}}^{(1)}_{t_p+1,i_0}\|\|\widetilde{\boldsymbol{\xi}}^{(2)}_{t_p+1,i_0}\|\lambda_{\max}(\boldsymbol{P}^{-1}_{t_p+1,i_0})\\ &= O\Bigg(\lambda_{\max}(\boldsymbol{P}^{-1}_{t_p+1,i_0})\Bigg[\frac{\alpha_{t_p+1,i_0}}{\lambda_{\min}(\boldsymbol{P}^{-1}_{t_p+1,i_0})}\\ &\qquad + \sqrt{\frac{\log(r_{t_p})}{\lambda_{\min}(\boldsymbol{P}^{-1}_{t_p+1,i_0})}}\Bigg]\|\widetilde{\boldsymbol{\xi}}^{(2)}_{t_p+1,i_0}\|\Bigg),\end{aligned} \tag{31}$$

where c_3 and c_4 are two positive constants. So does $I^{(4)}_{t_p+1,i_0}$. Then by the definition of $\hat{\boldsymbol{\theta}}_{t_p+1,i_0}(l)$ in (15), and the condition $\log r_t = O(\log r_{t-D_{\mathcal{G}}+1})$, we have for $l = d+1,\ldots,m$, $L_{t_p+1,i_0} \leq |\hat{\boldsymbol{\theta}}_{t_p+1,i_0}(l)| \leq c_5 L_{t_p+1,i_0}$, where $c_5 > 0$ is a positive constant, and $L_{t_p+1,i_0} = \sqrt{\frac{\log(\lambda_{\max}(\boldsymbol{P}^{-1}_{t_p+1,i_0}))}{\lambda_{\min}(\boldsymbol{P}^{-1}_{t_p+1,i_0})}}$. Hence we have

$$\begin{aligned}I^{(5)}_{t_p+1,i_0} &\geq \alpha_{t_p+1,i_0}\frac{1}{c_5 L_{t_p+1,i_0}}\sum_{l=d+1}^{m}|\widetilde{\boldsymbol{\xi}}_{t_p+1,i_0}(l)|\\ &\geq \alpha_{t_p+1,i_0}\frac{1}{c_5 L_{t_p+1,i_0}}\|\widetilde{\boldsymbol{\xi}}^{(2)}_{t_p+1,i_0}\|.\end{aligned} \tag{32}$$

Thus, by (28)–(32), for some $c_6 > 0$, we obtain

$$\bar{J}_{t_p+1,i_0}(\boldsymbol{\theta}+\widetilde{\boldsymbol{\xi}}_{t_p+1,i_0}) - \bar{J}_{t_p+1,i_0}(\boldsymbol{\theta}+\bar{\boldsymbol{\xi}}_{t_p+1,i_0})$$

$$\geq \lambda_{\min}(\boldsymbol{P}^{-1}_{t_p+1,i_0})\|\widetilde{\boldsymbol{\xi}}^{(2)}_{t_p+1,i_0}\| \cdot \Bigg(\frac{\|\widetilde{\boldsymbol{\xi}}^{(2)}_{t_p+1,i_0}\|}{2} - c_2\sqrt{\frac{\lambda_{\max}(\boldsymbol{P}^{-1}_{t_p+1,i_0})}{\lambda_{\min}(\boldsymbol{P}^{-1}_{t_p+1,i_0})}}\sqrt{\frac{\log r_{t_p}}{\lambda_{\min}(\boldsymbol{P}^{-1}_{t_p+1,i_0})}} - \frac{c_6\lambda_{\max}(\boldsymbol{P}^{-1}_{t_p+1,i_0})}{\lambda_{\min}(\boldsymbol{P}^{-1}_{t_p+1,i_0})}\Bigg[\frac{\alpha_{t_p+1,i_0}}{\lambda_{\min}(\boldsymbol{P}^{-1}_{t_p+1,i_0})} + \sqrt{\frac{\log(r_{t_p})}{\lambda_{\min}(\boldsymbol{P}^{-1}_{t_p+1,i_0})}}\Bigg] + \frac{\alpha_{t_p+1,i_0}}{c_5\lambda_{\min}(\boldsymbol{P}^{-1}_{t_p+1,i_0})L_{t_p+1,i_0}}\Bigg). \tag{33}$$

By (13), (16) and Assumption 3, we have

$$\frac{\lambda_{\max}(\boldsymbol{P}^{-1}_{t_p+1,i_0})}{\lambda_{\min}(\boldsymbol{P}^{-1}_{t_p+1,i_0})}\sqrt{\frac{\log r_{t_p}}{\lambda_{\min}(\boldsymbol{P}^{-1}_{t_p+1,i_0})}} \leq \frac{\lambda_{\max}(\boldsymbol{P}^{-1}_{t_p+1,i_0})}{\lambda_{\min}(\boldsymbol{P}^{-1}_{t_p+1,i_0})}\sqrt{\frac{\log r_{t_p}}{\lambda^{n,t_p}_{\min}}} = o\Big(\frac{\alpha_{t_p+1,i_0}}{\lambda_{\min}(\boldsymbol{P}^{-1}_{t_p+1,i_0})L_{t_p+1,i_0}}\Big). \tag{34}$$

By (13) and Assumption 3, we have

$$\frac{\lambda_{\max}(\boldsymbol{P}^{-1}_{t_p+1,i_0})\alpha_{t_p+1,i_0}}{\lambda^2_{\min}(\boldsymbol{P}^{-1}_{t_p+1,i_0})}\Big/\frac{\alpha_{t_p+1,i_0}}{\lambda_{\min}(\boldsymbol{P}^{-1}_{t_p+1,i_0})L_{t_p+1,i_0}} = L_{t_p+1,i_0}\frac{\lambda_{\max}(\boldsymbol{P}^{-1}_{t_p+1,i_0})}{\lambda_{\min}(\boldsymbol{P}^{-1}_{t_p+1,i_0})} = O\Big(\frac{r_{t_p}}{\lambda^{n,t_p}_{\min}}\sqrt{\frac{\log(r_{t_p})}{\lambda^{n,t_p}_{\min}}}\Big) = o(1). \tag{35}$$

From (33)–(35), we have

$$\bar{J}_{t_p+1,i_0}(\boldsymbol{\theta} + \widetilde{\boldsymbol{\xi}}_{t_p+1,i_0}) - \bar{J}_{t_p+1,i_0}(\boldsymbol{\theta} + \bar{\boldsymbol{\xi}}_{t_p+1,i_0}) \geq \lambda_{\min}(\boldsymbol{P}^{-1}_{t_p+1,i_0})\|\widetilde{\boldsymbol{\xi}}^{(2)}_{t_p+1,i_0}\| \cdot \Bigg(\frac{\|\widetilde{\boldsymbol{\xi}}^{(2)}_{t_p+1,i_0}\|}{2} + \frac{[\frac{1}{c_5} + o(1)]\alpha_{t_p+1,i_0}}{\lambda_{\min}(\boldsymbol{P}^{-1}_{t_p+1,i_0})L_{t_p+1,i_0}}\Bigg). \tag{36}$$

Note that $\widetilde{\boldsymbol{\xi}}_{t_p+1,i_0}(s_l) \neq 0$ for some $s_l \in \{d+1, \ldots, m\}$. Hence $\|\widetilde{\boldsymbol{\xi}}^{(2)}_{t_p+1,i_0}\| > 0$. Then by (36), we have $J_{t_p+1,i_0}(\boldsymbol{\theta} + \widetilde{\boldsymbol{\xi}}_{t_p+1,i_0}) - \bar{J}_{t_p+1,i_0}(\boldsymbol{\theta} + \bar{\boldsymbol{\xi}}_{t_p+1,i_0}) > 0$, which contradicts (25). This implies that $\|\widetilde{\boldsymbol{\xi}}^{(2)}_{t+1,i}\| = 0$ for all large t and all $i \in \{1, \ldots, n\}$, which completes the proof. □

5. A simulation example

In this section, we provide an example to illustrate the performance of the distributed sparse identification algorithm (i.e., Algorithm 1) proposed in this paper.

Consider a network composed of $n = 18$ sensors whose dynamics obey the model (1) with the dimension $m = 16$. The noise sequence $\{w_{t,i}, t \geq 1, i = 1, \ldots, n\}$ in (1) is independent and identically distributed with $w_{t,i} \sim \mathcal{N}(0, 1)$ (Gaussian distribution with zero mean and variance 1). The regression vectors $\{\boldsymbol{\varphi}_{t,i} \in \mathbb{R}^{16}, i = 1, \ldots, 18, t \geq 1\}$ are generated according to the following expression.

$$\boldsymbol{\varphi}_{t,i} = \Big[0, \ldots, 0, \underbrace{1.2^t + \sum_{k=0}^{t-1}\sin(ik)\varepsilon_{t-k,i}}_{j^{th}}, 0, \ldots, 0\Big]^T,$$

where $j = \mathrm{mod}(i, m)$ and the noise sequences $\{\varepsilon_{t,i}, i = 1, \ldots, n, t \geq 1\}$ are independent and uniformly distributed in $(-1, 1)$. All sensors will estimate an unknown parameter

$$\boldsymbol{\theta} = [0, 0, 0.6, 1.5, 0, 0, 2, 0, 0, 0, 0.9, -1.2, 0, 1.8, 0, -0.6]^T.$$

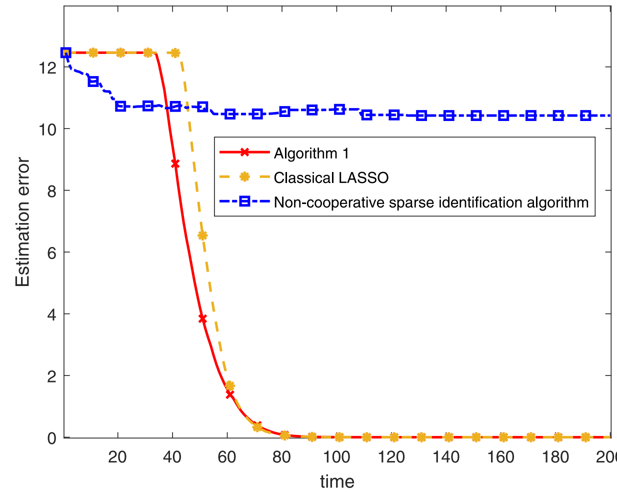

Fig. 1. The estimation errors of Algorithm 1, classical LASSO and non-cooperative sparse identification algorithm.

The initial estimate is taken as $\boldsymbol{\xi}_{0,i} = [1, \ldots, 1]^T$ for $i = 1, 2, \ldots, n$. We use the Metropolis rule (Xiao et al., 2005) to construct the weights of the network.

It can be verified that for each sensor i ($i = 1 \cdots, 18$), the regression signals $\boldsymbol{\varphi}_{t,i}$ have no adequate excitation to estimate the unknown parameter, but they can cooperate to satisfy Assumption 3.

(1) We estimate the unknown parameter $\boldsymbol{\theta}$ by using the non-cooperative sparse identification algorithm (i.e., the adjacency matrix is the unit matrix), the classical LASSO (Eq. (9)) and the distributed sparse identification algorithm (Algorithm 1) proposed in this paper respectively. We adopt the Matlab CVX tools to solve the convex optimization problem (17), and take the regularization coefficient as $\alpha_{t,i} = (\lambda_{\min}(\boldsymbol{P}^{-1}_{t+1,i}))^{0.75}$. The average estimation error generated by these three algorithms is shown in Fig. 1. We see that the estimation error generated by distributed sparse identification algorithm converges to zero as t increases, while the estimation error of the non-cooperative sparse identification algorithm does not. Therefore, the estimation task can be fulfilled through exchanging information between sensors even though any individual sensor cannot. Moreover, from Fig. 1, we can see that Algorithm 1 has better estimation performance than the classical LASSO.

(2) We estimate the unknown parameter $\boldsymbol{\theta}$ by using the classical distributed LS algorithm studied by Xie et al. (2021) and Algorithm 1 proposed in this paper under the same network topology. Table 1 shows the estimates for $\boldsymbol{\theta}(1)$ by these two algorithms at time instants $t = 50$ (similar for other zero elements in $\boldsymbol{\theta}$). From Table 1, we can see that, compared with the distributed LS algorithm in Xie et al. (2021), Algorithm 1 can generate sparser and more accurate estimates for the unknown parameters and thus give us valuable information in inferring the zero and nonzero elements in the unknown parameters.

6. Concluding remarks

In this paper, we introduced a local information criterion which is formulated as a linear combination of the local estimation error with L_1-regularization term. By minimizing this criterion, we proposed a distributed sparse identification algorithm to estimate an unknown parameter vector of a stochastic system. Under a cooperative non-persistent excitation condition, the set of zero elements in the unknown parameter vector can be correctly identified with a finite number of observations by properly

Table 1
Estimates for $\boldsymbol{\theta}(1)$ by the distributed LS algorithm in Xie et al. (2021) and Algorithm 1 at $t = 50$.

	Sensor 1	Sensor 2	Sensor 3	Sensor 4	Sensor 5	Sensor 6
By distributed LS	-3.82×10^{-4}	-3.84×10^{-4}	-2.53×10^{-4}	-1.29×10^{-4}	-3.09×10^{-4}	-1.92×10^{-4}
By Algorithm 1	3.18×10^{-13}	3.34×10^{-14}	2.31×10^{-14}	3.50×10^{-14}	2.99×10^{-14}	3.38×10^{-13}
	Sensor 7	Sensor 8	Sensor 9	Sensor 10	Sensor 11	Sensor 12
By distributed LS	-3.17×10^{-4}	-2.87×10^{-4}	-3.52×10^{-4}	-2.72×10^{-4}	-1.69×10^{-4}	-2.34×10^{-4}
By Algorithm 1	3.00×10^{-13}	1.14×10^{-13}	1.09×10^{-12}	3.30×10^{-13}	2.29×10^{-13}	2.17×10^{-13}
	Sensor 13	Sensor 14	Sensor 15	Sensor 16	Sensor 17	Sensor 18
By distributed LS	-3.31×10^{-4}	-3.86×10^{-4}	-3.12×10^{-4}	-4.52×10^{-5}	-4.24×10^{-5}	-3.68×10^{-4}
By Algorithm 1	1.75×10^{-13}	3.79×10^{-13}	7.92×10^{-14}	2.66×10^{-13}	4.01×10^{-13}	1.68×10^{-13}

choosing the weighting coefficient. We remark that our theoretical results are established without using such stringent conditions as independency of the regression vectors, which makes it possible to combine the distributed adaptive estimation with the distributed control. For future research, it will be interesting to consider the combination of the distributed sparse identification algorithm with the distributed control, and design a recursive distributed sparse adaptive algorithm.

References

Abdolee, R., & Champagne, B. (2016). Diffusion LMS strategies in sensor networks with noisy input data. *IEEE/ACM Transactions on Networking*, *24*(1), 3–14.

Baraniuk, R. G. (2007). Compressive sensing. *IEEE Signal Processing Magazine*, *24*(4), 118–121.

Battilotti, S., Cacace, F., d'Angelo, M., & Germani, A. (2020). Asymptotically optimal consensus-based distributed filtering of continuous-time linear systems. *Automatica*, *122*, Article 109189.

Bazerque, J. A., & Giannakis, G. B. (2010). Distributed spectrum sensing for cognitive radio networks by exploiting sparsity. *IEEE Transactions on Signal Processing*, *58*(3), 1847–1862.

Candès, E. J., & Tao, T. (2005). Decoding by linear programming. *IEEE Transactions on Information Theory*, *51*(12), 4203–4215.

Chen, Y., Gu, Y., & Hero, A. O. (2009). Sparse LMS for system identification. In *2009 IEEE international conference on acoustics, speech and signal processing* (pp. 3125–3128).

Chen, W., Hua, S., & Zhang, H. (2015). Consensus-based distributed cooperative learning from closed-loop neural control systems. *IEEE Transactions on Neural Networks and Learning Systems*, *26*, 331–345.

Chen, W., Wen, C., Hua, S., & Sun, C. (2014). Distributed cooperative adaptive identification and control for a group of continuous-time systems with a cooperative PE condition via consensus. *IEEE Transactions on Automatic Control*, *59*(1), 91–106.

Chiuso, A., & Pillonetto, G. (2014). Bayesian and nonparametric methods for system identification and model selection. In *2014 european control conference* (pp. 2376–2381).

Di Lorenzo, P., & Sayed, A. H. (2013). Sparse distributed learning based on diffusion adaptation. *IEEE Transactions on Signal Processing*, *61*(6), 1419–1433.

Eksioglu, E. M. (2013). Group sparse RLS algorithms. *International Journal of Adaptive Control and Signal Processing*, *28*(12), 1398–1412.

Gan, D., & Liu, Z. (2019). Strong consistency of the distributed stochastic gradient algorithm. In *Proceedings of the 58th IEEE conference on decision and control* (pp. 5082–5087). Nice, France.

Gan, D., & Liu, Z. (2022a). Distributed order estimation of ARX model under cooperative excitation condition. *SIAM Journal on Control and Optimization*, *60*(3), 1519–1545.

Gan, D., & Liu, Z. (2022b). Performance analysis of the compressed distributed least squares algorithm. *Systems & Conrol Letters*, *164*, Article 105228.

Gan, D., Xie, S., & Liu, Z. (2021). Stability of the distributed Kalman filter using general random coefficients. *Science China. Information Sciences*, *64*, Article 172204.

Gill, P. R., Wang, A., & Molnar, A. (2011). The in-crowd algorithm for fast basis pursuit denoising. *IEEE Transactions on Signal Processing*, *59*(10), 4595–4605.

Huang, W., Chen, C., Yao, X., & Li, Q. (2020). Diffusion fused sparse LMS algorithm over networks. *Signal Processing*, *171*, Article 107497.

Huang, S., & Li, C. (2015). Distributed sparse total least-squares over networks. *IEEE Transactions on Signal Processing*, *63*(11), 2986–2998.

Kaiser, E., Kutz, J. N., & L., B. S. (2018). Sparse identification of nonlinear dynamics for model predictive control in the low-data limit. *Proceedings of the Royal Society of London, Series A (Mathematical and Physical Sciences)*, *474*(2219), Article 20180335.

Kim, S.-J., Koh, K., Lustig, M., Boyd, S., & Gorinevsky, D. (2007). An interior-point method for large-scale ℓ_1-regularized least squares. *IEEE Journal of Selected Topics in Signal Processing*, *1*(4), 606–617.

Liu, Y., Liu, J., Xu, C., Li, G., & He, Y. (2020). Fully distributed variational Bayesian non-linear filter with unknown measurement noise in sensor networks. *Science China. Information Sciences*, *63*, Article 210202.

Sayed, A. H., Tu, S.-Y., Chen, J., Zhao, X., & Towfic, Z. J. (2013). Diffusion strategies for adaptation and learning over networks: an examination of distributed strategies and network behavior. *IEEE Signal Processing Magazine*, *30*(3), 155–171.

Shiri, H., Tinati, M. A., Codreanu, M., & Daneshvar, S. (2018). Distributed sparse diffusion estimation based on set membership and affine projection algorithm. *Digital Signal Processing*, *73*, 47–61.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *58*(1), 267–288.

Vinga, S. (2021). Structured sparsity regularization for analyzing high-dimensional omics data. *Brief Bioinform*, *22*(1), 77–87.

Xiao, L., Boyd, S., & Lall, S. (2005). A scheme for robust distributed sensor fusion based on average consensus. In *Proceedings of the 4th fourth international symposium on information processing in sensor networks* (pp. 63–70). Boise, ID, USA.

Xie, S., & Guo, L. (2018). A necessary and sufficient condition for stability of LMS-based consensus adaptive filters. *Automatica*, *93*, 12–19.

Xie, S., & Guo, L. (2020). Analysis of compressed distributed adaptive filters. *Automatica*, *112*, Article 108707.

Xie, S., Zhang, Y., & Guo, L. (2021). Convergence of a distributed least squares. *IEEE Transactions on Automatic Control*, *66*(10), 4952–4959.

Xu, S., de Lamare, R. C., & Poor, H. V. (2015). Distributed compressed estimation based on compressive sensing. *IEEE Signal Processing Letters*, *22*(9), 1311–1315.

Zhang, H., Wang, T., & Zhao, Y. (2021). Asymptotically efficient recursive identification of FIR systems with binary-valued observations. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *51*(5), 2687–2700.

Zhao, W., Yin, G., & Bai, E.-W. (2020). Sparse system identification for stochastic systems with general observation sequences. *Automatica*, *121*, Article 109162.

Zhao, P., & Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research*, *7*, 2541–2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429.

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 postdoctoral fellow with the Zhongguancun Laboratory, Beijing, China. She received China National Scholarship for graduate students in 2020, and the outstanding graduate title of Beijing in 2022. Her research interests include system identification, machine learning and distributed adaptive filter.

Zhixin Liu 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, and the director of Key Laboratory of Systems and Control, CAS. She is a coauthor of the SIGEST Paper in SIREV in 2014. Her current research interests include Multiagent systems, distributed estimation and control.