Convergence of a Distributed Least Squares

Siyu Xie, Yaqi Zhang, and Lei Guo, *Fellow, IEEE*

***Abstract*—In this paper, we consider a least-squares (LS)-based distributed algorithm build on a sensor network to estimate an unknown parameter vector of a dynamical system, where each sensor in the network has partial information only but is allowed to communicate with its neighbors. Our main task is to generalize the well-known theoretical results on the traditional LS to the current distributed case by establishing both the upper bound of the accumulated regrets of the adaptive predictor and the convergence of the distributed LS estimator, with the following key features compared with the existing literature on distributed estimation: Firstly, our theory does not need the previously imposed independence, stationarity or Gaussian property on the system signals, and hence is applicable to stochastic systems with feedback. Secondly, the cooperative excitation condition introduced and used in this paper for the convergence of the distributed LS estimate is the weakest possible one, which shows that even if any individual sensor cannot estimate the unknown parameter by the traditional LS, the whole network can still fulfill the estimation task by the distributed LS.**

***Index Terms*—Least squares, distributed estimation, convergence, diffusion strategies, cooperative excitation, regret, martingale theory**

I. Introduction

Distributed estimation over sensor networks has received increasing research attention recently, and has been studied and used in many areas widely, e.g., collaborative spectral sensing in cognitive radio systems, target localization in biological networks, environmental monitoring, military surveillance, and so on [1]. Note that different cooperation strategies will lead to different distributed estimation algorithms, for examples, incremental least mean squres (LMS) [2], consensus LMS [3], [4], diffusion LMS [5], [6], incremental LS [7], consensus LS [8], diffusion LS [9]–[15], and distributed Kalman filter (KF) [16]–[19]. In our recent work (see e.g. [3]–[5]), we have given the stability and performance results for the consensus and diffusion LMS filters, without imposing the usual independence and stationarity assumptions for the system signals.

Note that when the unknown parameter is time-invariant, the LS algorithm may generate more accurate estimates in the transient phase and have faster convergence speed compared with LMS algorithm. This is one of the main motivations for us to consider the LS-based distributed estimation algorithm in this paper. Another reason for us to study this problem is that the existing convergence theory in the literature [7]–[15] can hardly be applied to non-independent and non-stationary signals coming from practical complex systems where feedback loops exist inevitably.

Fortunately, in the traditional single sensor case, there is a vast literature on the convergence theory of the classical LS, which is indeed applicable to stochastic systems with feedback. In fact, motivated by the need to establish a rigorous theory for the well-known LS-based self-tuning regulators proposed by Åström and Wittenmark [20] in stochastic adaptive control, the convergence study of LS with possible stochastic feedback signals had received a great deal of attention in the literature, see e.g., [21]–[29]. At the same time, much effort had also been devoted to stochastic adaptive control, see e.g, [30]–[32]. Among the many significant contributions in this direction, here we only mention that Lai and Wei [25] established a celebrated convergence result under a weakest possible decaying excitation condition on the system signals, and Guo and Chen [29] and Guo [21] finally resolved the longstanding problem concerning the global stability and convergence of the LS-based self-tuning regulators.

In this paper, we will provide a theoretical analysis for a distributed LS algorithm of diffusion type [17], [18], where the diffusion strategy is designed via the so called covariance intersection fusion rule [33], [34]. In such a diffusion strategy, each node is only allowed to communicate with its neighbors, and both the estimates of the unknown parameter and the inverse of the covariance matrices are diffused between neighboring nodes. We will generalize the well-known convergence results on the classical LS by establishing both the upper bound of the accumulated regrets of the adaptive predictor and the convergence of the distributed LS estimator, with the following key features compared with the related results in the existing literature:

- Our theory does not need the usually assumed independence, stationarity or Gaussian property on the system signals, and hence does not exclude the applications of the theory to stochastic feedback systems.
- Our theory for the convergence of the distributed LS is established under a weakest possible cooperative excitation condition which is a natural extension of the single sensor case. The cooperative excitation condition introduced in this paper implies that even if any individual sensor is not able to estimate the unknown parameter, the distributed LS can still accomplish the estimation task.

The rest of the paper is organized as follows. In Section II, we present some preliminaries on notations and graph theory, the observation model, and the distributed LS algorithm studied in the paper. The main results are stated in Section III. In Section IV, we provide the proofs of the main results. Finally, some concluding remarks are given in Section V.

S. Y. Xie is with the Department of Electrical and Computer Engineering, Wayne State University, Detroit, MI 48202, USA. Email: syxie@wayne.edu.

Y. Q. Zhang and L. Guo are with Key Laboratory of Systems and Control, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, P. R. China. They are also with School of Mathematical Science, University of Chinese Academy of Sciences, Beijing 100049, P. R. China. Email : *zhangyq@amss.ac.cn., lguo@amss.ac.cn.*

This work was supported by the National Natural Science Foundation of China under grant 11688101.

II. Problem Formulation

A. Basic Notations

Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times n}$ be two symmetric matrices with real entries, then $A \geq B (A > B)$ means $A - B$ is a positive semidefinite (definite) matrix. Also, let $\lambda_{\max}\{\cdot\}$ and $\lambda_{\min}\{\cdot\}$ denote the largest and the smallest eigenvalues of the corresponding matrix, respectively. For any matrix $X \in \mathbb{R}^{m\times n}$, $\| X \|$ denotes the operator norm induced by the Euclidean norm, i.e., $(\lambda_{\max}\{XX^\top\})^{\frac{1}{2}}$, where $(\cdot)^\top$ denotes the transpose operator. We use $\mathbb{E}[\cdot]$ to denote the mathematical expectation operator, and $\mathbb{E}[\cdot|\mathcal{F}_k]$ to denote the conditional mathematical expectation operator, where $\{\mathcal{F}_k\}$ is a sequence of nondecreasing σ-algebras [35]. We also use $\log(\cdot)$ to denote the natural logarithm function, and $\mathrm{Tr}(\cdot)$ to denote the trace of the corresponding matrix. Through out the paper, $|\cdot|$ denotes the determinant of the corresponding matrix, which should not be confused with the absolute value of a scalar from the context.

Let $\{A_k, k \geq 0\}$ be a matrix sequence and $\{b_k, k \geq 0\}$ be a positive scalar sequence. Then by $A_k = O(b_k)$ we mean that there exists a constant $M > 0$ such that $\|A_k\| \leq Mb_k, \forall k \geq 0$, and by $A_k = o(b_k)$ we mean that $\lim\limits_{k\to\infty} \|A_k\|/b_k = 0$.

B. Graph Theory

As usual, let the communication structure among sensors be represented by an undirected weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$, where $\mathcal{V} = \{1, 2, ..., n\}$ is the set of sensors and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges. The structure of the graph $\mathcal{G}$ is described by $\mathcal{A} = \{a_{ij}\}_{n\times n}$ which is called the weighted adjacency matrix, where $a_{ij} > 0$ if $(i, j) \in \mathcal{E}$ and $a_{ij} = 0$ otherwise. In this paper, we assume that the elements of the weighted matrix $\mathcal{A}$ satisfy $a_{ij} = a_{ji}, \forall i, j = 1, \ldots, n$, and $\sum_{j=1}^n a_{ij} = 1, \forall i = 1, \ldots, n$. Thus the matrix $\mathcal{A}$ is symmetric and doubly stochastic[1].

A path of length ℓ in the graph $\mathcal{G}$ is a sequence of nodes $\{i_1, \ldots, i_\ell\}$ subject to $(i_j, i_{j+1}) \in \mathcal{E}$, for $1 \leq j \leq \ell - 1$. The maximum value of the distances between any two nodes in the graph $\mathcal{G}$ is called the diameter of $\mathcal{G}$. Here in this paper, we assume that the graph is connected, and denote the diameter of the graph $\mathcal{G}$ as $D_\mathcal{G}$. Then $1 \leq D_\mathcal{G} < \infty$ holds. The set of neighbors of the sensor i is denoted as $\mathcal{N}_i = \{j \in V|(j, i) \in \mathcal{E}\}$, and the sensor i can only share information with its neighboring sensors from $\mathcal{N}_i$.

C. Observation Model

Let us consider a sensor network consisting of n sensors. Assume that at each time instant k, each sensor $i \in \{1, \ldots, n\}$ in the sensor network receives a noisy scalar measurement $y_{k+1,i}$ and an m-dimensional regressor $\boldsymbol{\varphi}_{k,i} \in \mathbb{R}^m$. They are related by a typical linear stochastic regression model

$$y_{k+1,i} = \boldsymbol{\varphi}_{k,i}^\top \boldsymbol{\theta} + w_{k+1,i}, \qquad k \geq 0, \tag{1}$$

where $w_{k+1,i}$ is a random noise process, and $\boldsymbol{\theta} \in \mathbb{R}^m$ is an unknown parameter vector which needs to be estimated. We assume that at any time instant $k \geq 1$, each sensor i uses both the observations $y_{j+1,i}$ and the regressors $\boldsymbol{\varphi}_{j,i} (j \leq k)$ to estimate the unknown parameter $\boldsymbol{\theta}$.

[1]A matrix is called doubly stochastic, if all elements are nonnegative, both the sum of each row and the sum of each column equal to 1.

D. Distributed LS Algorithm

We now consider the following basic class of distributed LS algorithms of diffusion type:

Algorithm 1 A Distributed LS algorithm

For any given sensor $i \in \{1 \ldots, n\}$, begin with an initial estimate $\boldsymbol{\theta}_{0,i} \in \mathbb{R}^m$, and an initial positive definite matrix $P_{0,i} \in \mathbb{R}^{m\times m}$. The algorithm is recursively defined at any time instant $k \geq 0$ as follows:

1: Adapt (generate $\bar{\boldsymbol{\theta}}_{k+1,i}$ and $\bar{P}_{k+1,i}$ on the bases of $\boldsymbol{\theta}_{k,i}, P_{k,i}$ and $\boldsymbol{\varphi}_{k,i}, y_{k+1,i}$):

$$\bar{\boldsymbol{\theta}}_{k+1,i} = \boldsymbol{\theta}_{k,i} + b_{k,i} P_{k,i} \boldsymbol{\varphi}_{k,i} (y_{k+1,i} - \boldsymbol{\varphi}_{k,i}^\top \boldsymbol{\theta}_{k,i}), \tag{2}$$

$$\bar{P}_{k+1,i} = P_{k,i} - b_{k,i} P_{k,i} \boldsymbol{\varphi}_{k,i} \boldsymbol{\varphi}_{k,i}^\top P_{k,i}, \tag{3}$$

$$b_{k,i} = (1 + \boldsymbol{\varphi}_{k,i}^\top P_{k,i} \boldsymbol{\varphi}_{k,i})^{-1}, \tag{4}$$

2: Combine (generate $P_{k+1,i}^{-1}$ and $\boldsymbol{\theta}_{k+1,i}$ by a convex combination of $\bar{P}_{k+1,j}^{-1}$ and $\bar{\boldsymbol{\theta}}_{k+1,j}$):

$$P_{k+1,i}^{-1} = \sum_{j\in\mathcal{N}_i} a_{ji} \bar{P}_{k+1,j}^{-1}, \tag{5}$$

$$\boldsymbol{\theta}_{k+1,i} = P_{k+1,i} \sum_{j\in\mathcal{N}_i} a_{ji} \bar{P}_{k+1,j}^{-1} \bar{\boldsymbol{\theta}}_{k+1,j}. \tag{6}$$

Remark 2.1: When $\mathcal{A} = I_n$, the distributed LS will degenerate to the classical LS at any sensor i. Note that for stochastic gradient-based [28] and LMS-based [36] distributed estimation algorithms, the communication complexity may be reduced. However, for those algorithms, the estimation error either converges slowly to zero or does not converge to zero at all. Therefore, there is a tradeoff between the complexity and the convergence rate of the distributed estimation algorithms. Moreover, the convergence rate would be "optimal" when $\bar{P}_{k,i}$ is chosen to be the form in the paper. Furthermore, some existing methods may be used to reduce the communication complexity and to make the algorithm suitable for higher dimensional signals, for examples, event-driven methods [37], partial diffusion methods [12], [13], and compressed methods [38] and so on.

III. The Main Results

A. Some Preliminaries

For the theoretical analysis, we need some standard assumptions on noise processes, regressors, and network topology.

Assumption 3.1: For each $i \in \{1, \ldots, n\}$, the noise sequence $\{w_{k,i}, \mathcal{F}_k\}$ is a martingale difference (where $\{\mathcal{F}_k\}$ is a sequence of nondecreasing σ-algebras), and there exists a constant $\beta > 2$ such that $\sup_{k\geq 0} \mathbb{E}[|w_{k+1,i}|^\beta|\mathcal{F}_k] < \infty$, a.s.

Assumption 3.2: For each $i \in \{1, \ldots, n\}$, the regressor sequence $\{\boldsymbol{\varphi}_{k,i}, \mathcal{F}_k\}$ is an adapted sequence.

Assumption 3.3: The graph $\mathcal{G}$ is connected.

Remark 3.1: From *Lemma 8.1.2* in [39], it is not difficult to see that for any two nodes i and j, there exists a path from i to j with length not larger than ℓ if and only if the (i,j)th entry of the matrix $\mathcal{A}^\ell$ is positive. From this, it is easy to see that each entry of the matrix $\mathcal{A}^\ell$ will be positive when ℓ is not smaller than the diameter of the graph $\mathcal{G}$, i.e., $D_{\mathcal{G}}$, see also [18].

B. Theoretical Results

Here we first introduce the following notations:

$$\begin{aligned}
&\boldsymbol{Y}_k \triangleq \text{col}\{y_{k,1},\ldots,y_{k,n}\},\quad \boldsymbol{\Phi}_k \triangleq \text{diag}\{\boldsymbol{\varphi}_{k,1},\ldots,\boldsymbol{\varphi}_{k,n}\},\\
&\boldsymbol{W}_k \triangleq \text{col}\{w_{k,1},\ldots,w_{k,n}\},\quad \boldsymbol{\Theta} \triangleq \text{col}\{\underbrace{\boldsymbol{\theta},\ldots,\boldsymbol{\theta}}_{n}\},\\
&\boldsymbol{\Theta}_k \triangleq \text{col}\{\boldsymbol{\theta}_{k,1},\ldots,\boldsymbol{\theta}_{k,n}\},\quad \bar{\boldsymbol{\Theta}}_k \triangleq \text{col}\{\bar{\boldsymbol{\theta}}_{k,1},\ldots,\bar{\boldsymbol{\theta}}_{k,n}\},\\
&\widetilde{\boldsymbol{\Theta}}_k \triangleq \text{col}\{\widetilde{\boldsymbol{\theta}}_{k,1},\ldots,\widetilde{\boldsymbol{\theta}}_{k,n}\},\quad \text{where } \widetilde{\boldsymbol{\theta}}_{k,i}=\boldsymbol{\theta}-\boldsymbol{\theta}_{k,i},\\
&\widetilde{\bar{\boldsymbol{\Theta}}}_k \triangleq \text{col}\{\widetilde{\bar{\boldsymbol{\theta}}}_{k,1},\ldots,\widetilde{\bar{\boldsymbol{\theta}}}_{k,n}\},\quad \text{where } \widetilde{\bar{\boldsymbol{\theta}}}_{k,i}=\boldsymbol{\theta}-\bar{\boldsymbol{\theta}}_{k,i},\\
&\boldsymbol{P}_k \triangleq \text{diag}\{P_{k,1},\ldots,P_{k,n}\},\quad \bar{\boldsymbol{P}}_k \triangleq \text{diag}\{\bar{P}_{k,1},\ldots,\bar{P}_{k,n}\},\\
&\boldsymbol{b}_k \triangleq \text{diag}\{b_{k,1},\ldots,b_{k,n}\},\quad \boldsymbol{c}_k \triangleq \boldsymbol{b}_k\otimes I_m,\quad \mathscr{A}\triangleq \mathcal{A}\otimes I_m,
\end{aligned}$$

where $\text{col}\{\cdots\}$ denotes a vector by stacking the specified vectors, $\text{diag}\{\cdots\}$ is used in a non-standard manner which means that $m\times 1$ column vectors are combined "in a diagonal manner" resulting in a $mn\times n$ matrix, and $\otimes$ is the Kronecker product. Then (1) can be rewritten in the following compact form:

$$\boldsymbol{Y}_{k+1}=\boldsymbol{\Phi}_k^\top\boldsymbol{\Theta}+\boldsymbol{W}_{k+1}, \tag{7}$$

Similarly, for the distributed LS algorithm we have

$$\left\{\begin{array}{l}
\bar{\boldsymbol{\Theta}}_{k+1}=\boldsymbol{\Theta}_k+\boldsymbol{c}_k\boldsymbol{P}_k\boldsymbol{\Phi}_k(\boldsymbol{Y}_{k+1}-\boldsymbol{\Phi}_k^\top\boldsymbol{\Theta}_k),\\
\bar{\boldsymbol{P}}_{k+1}=\boldsymbol{P}_k-\boldsymbol{c}_k\boldsymbol{P}_k\boldsymbol{\Phi}_k\boldsymbol{\Phi}_k^\top\boldsymbol{P}_k,\\
\boldsymbol{b}_k=(I_n+\boldsymbol{\Phi}_k^\top\boldsymbol{P}_k\boldsymbol{\Phi}_k)^{-1},\\
\boldsymbol{c}_k=\boldsymbol{b}_k\otimes I_m,\\
\text{vec}\{\boldsymbol{P}_{k+1}^{-1}\}=\mathscr{A}\text{vec}\{\bar{\boldsymbol{P}}_{k+1}^{-1}\},\\
\boldsymbol{\Theta}_{k+1}=\boldsymbol{P}_{k+1}\mathscr{A}\bar{\boldsymbol{P}}_{k+1}^{-1}\bar{\boldsymbol{\Theta}}_{k+1},
\end{array}\right. \tag{8}$$

where $\text{vec}\{\cdot\}$ denotes the operator that stacks the blocks of a block diagonal matrix on top of each other.

Since $\widetilde{\boldsymbol{\Theta}}_k=\boldsymbol{\Theta}-\boldsymbol{\Theta}_k$ and $\widetilde{\bar{\boldsymbol{\Theta}}}_k=\boldsymbol{\Theta}-\bar{\boldsymbol{\Theta}}_k$ by definition, substituting (7) into (8), we can get

$$\begin{aligned}
\widetilde{\bar{\boldsymbol{\Theta}}}_{k+1}=&\boldsymbol{\Theta}-\boldsymbol{\Theta}_k-\boldsymbol{c}_k\boldsymbol{P}_k\boldsymbol{\Phi}_k(\boldsymbol{\Phi}_k^\top\boldsymbol{\Theta}+\boldsymbol{W}_{k+1}-\boldsymbol{\Phi}_k^\top\boldsymbol{\Theta}_k)\\
=&(I_{mn}-\boldsymbol{c}_k\boldsymbol{P}_k\boldsymbol{\Phi}_k\boldsymbol{\Phi}_k^\top)\widetilde{\boldsymbol{\Theta}}_k-\boldsymbol{c}_k\boldsymbol{P}_k\boldsymbol{\Phi}_k\boldsymbol{W}_{k+1}\\
=&\bar{\boldsymbol{P}}_{k+1}\boldsymbol{P}_k^{-1}\widetilde{\boldsymbol{\Theta}}_k-\boldsymbol{c}_k\boldsymbol{P}_k\boldsymbol{\Phi}_k\boldsymbol{W}_{k+1}.
\end{aligned}$$

Note also that

$$\begin{aligned}
&\boldsymbol{P}_{k+1}\mathscr{A}\bar{\boldsymbol{P}}_{k+1}^{-1}\boldsymbol{\Theta}\\
=&\text{col}\left\{P_{k+1,1}\sum_{j\in\mathcal{N}_1}a_{j1}\bar{P}_{k+1,j}^{-1}\boldsymbol{\theta},\ldots,P_{k+1,n}\sum_{j\in\mathcal{N}_n}a_{jn}\bar{P}_{k+1,j}^{-1}\boldsymbol{\theta}\right\}\\
=&\boldsymbol{\Theta}.
\end{aligned}$$

Then we have

$$\begin{aligned}
\widetilde{\boldsymbol{\Theta}}_{k+1}=&\boldsymbol{\Theta}-\boldsymbol{P}_{k+1}\mathscr{A}\bar{\boldsymbol{P}}_{k+1}^{-1}\bar{\boldsymbol{\Theta}}_{k+1}=\boldsymbol{P}_{k+1}\mathscr{A}\bar{\boldsymbol{P}}_{k+1}^{-1}\widetilde{\bar{\boldsymbol{\Theta}}}_{k+1}\\
=&\boldsymbol{P}_{k+1}\mathscr{A}\boldsymbol{P}_k^{-1}\widetilde{\boldsymbol{\Theta}}_k\\
&-\boldsymbol{P}_{k+1}\mathscr{A}\bar{\boldsymbol{P}}_{k+1}^{-1}\boldsymbol{c}_k\boldsymbol{P}_k\boldsymbol{\Phi}_k\boldsymbol{W}_{k+1}.
\end{aligned} \tag{9}$$

Before establishing the convergence of the distributed LS, we first present a critical theorem, which requires no excitation conditions on the regression process $\boldsymbol{\varphi}_{k,i}$.

Theorem 3.1: Let *Assumptions 3.1* and *3.2* be satisfied, we have as $t\to\infty$,

1) $\sum_{k=0}^{t}\widetilde{\boldsymbol{\Theta}}_k^\top\boldsymbol{\Phi}_k\boldsymbol{b}_k\boldsymbol{\Phi}_k^\top\widetilde{\boldsymbol{\Theta}}_k=O(\log(r_t)),\quad$ a.s.,
2) $\widetilde{\boldsymbol{\Theta}}_{t+1}^\top\boldsymbol{P}_{t+1}^{-1}\widetilde{\boldsymbol{\Theta}}_{t+1}=O(\log(r_t)),\quad$ a.s.,

where

$$r_t=\lambda_{\max}\{\boldsymbol{P}_0^{-1}\}+\sum_{i=1}^{n}\sum_{k=0}^{t}\|\boldsymbol{\varphi}_{k,i}\|^2. \tag{10}$$

From this, we can obtain the upper bound of the accumulated regrets for the distributed LS-based adaptive predictor. For any $i\in\{1,\ldots,n\}$, and at any time instant $k\geq 1$, the best prediction to the future observation $y_{k+1,i}$ is the conditional mathematical expectation $\mathbb{E}[y_{k+1,i}|\mathcal{F}_k]=\boldsymbol{\varphi}_{k,i}^\top\boldsymbol{\theta}$, since the noise is a martingale difference sequence with second moment. Unfortunately, this optimal predictor is unavailable because $\boldsymbol{\theta}$ is unknown. A natural way is to construct an adaptive predictor $\widehat{y}_{k+1,i}$ by using the online distributed LS estimate $\boldsymbol{\theta}_{k,i}$, i.e., $\widehat{y}_{k+1,i}=\boldsymbol{\varphi}_{k,i}^\top\boldsymbol{\theta}_{k,i}$. The error between the best predictor and the adaptive predictor is referred to as the regret denoted by

$$R_{k,i}=(\mathbb{E}[y_{k+1,i}|\mathcal{F}_k]-\widehat{y}_{k+1,i})^2, \tag{11}$$

which may not be zero and even may not be small in sample paths due to the persistent disturbance of the unpredictable noises in the model.

However, one may evaluate the averaged regrets defined as follows:

$$\frac{1}{nt}\sum_{i=1}^{n}\sum_{k=0}^{t}R_{k,i},$$

which we are going to show tends to zero as t increases to infinity under essentially no excitation conditions on the regressors, see *Theorem 3.2* below. This is a celebrated property that has been widely studied in distributed online learning and optimization problems [40], [41], but under rather restrictive assumptions such as boundedness, stationarity or independence on the system signals.

Theorem 3.2: Let *Assumptions 3.1* and *3.2* be satisfied. Then the sample paths of the accumulated regrets have the following bound as $t\to\infty$:

$$\sum_{i=1}^{n}\sum_{k=0}^{t}R_{k,i}=O(\log(r_t)),\qquad\text{a.s.,} \tag{12}$$

provided that $\boldsymbol{\Phi}_t^\top\boldsymbol{P}_t\boldsymbol{\Phi}_t=O(1)$, a.s.

Remark 3.2: We remark that the order $O(\log(r_t))$ for the accumulated regrets may be shown to be the best possible among all adaptive predictors in a certain sense, as is already known in the traditional single sensor case, see [42]. The precise constant in $O(\cdot)$ may also be determined if we have

further conditions on the regressors, see *Corollary 3.3* in [21] in the single sensor case.

From *Theorem 3.1*, we can also obtain the strong consistency of the distributed LS to guarantee the generalization ability of learning.

Theorem 3.3: Let *Assumptions 3.1-3.3* be satisfied, we have as $t \to \infty$,

$$\|\widetilde{\boldsymbol{\Theta}}_{t+1}\|^2 = O\left(\frac{\log(r_t)}{\lambda_{\min}^{n,t}}\right), \quad \text{a.s.,} \tag{13}$$

where r_t is defined by (10) and

$$\lambda_{\min}^{n,t} = \lambda_{\min}\left\{\sum_{j=1}^{n} P_{0,j}^{-1} + \sum_{j=1}^{n}\sum_{k=0}^{t-D_{\mathcal{G}}+1} \boldsymbol{\varphi}_{k,j}\boldsymbol{\varphi}_{k,j}^{\top}\right\}. \tag{14}$$

Remark 3.3: Theorem 3.3 shows that if

$$\lim_{t\to\infty} \frac{\log(r_t)}{\lambda_{\min}^{n,t}} = 0, \qquad \text{a.s.,} \tag{15}$$

then the distributed LS estimate $\boldsymbol{\Theta}_t$ will converge to the true unknown parameter. We may name (15) as a cooperative excitation condition. In the traditional single sensor case (where $n = 1$, $D_{\mathcal{G}} = 1$), (15) reduces to the well-known Lai-Wei excitation condition, which is known to be the weakest possible data condition for the convergence of the classical LS estimates [25], and is much weaker than the well-known persistence of excitation (PE) condition usually used in the parameter estimation of finite-dimensional linear control systems.

Moreover, it is easy to convince oneself that the cooperative excitation condition (15) will make it possible for the distributed LS to consistently estimate the unknown parameter, even if any individual sensor cannot due to lack of suitable excitation, thanks to the cooperative nature of the excitation condition (15). Finally, we remark that the verification of (15) is straightforward in the ergodic case. For more general correlated non-stationary signals from control systems, the verification of (15) may be conducted following a similar way as that for the traditional single sensor case (see, [28]).

Furthermore, the convergence rate established in *Theorem 3.3* is essentially in terms of the increase of the number of observations rather than the number of iterations in computation.

Remark 3.4: Let us now compare the above distributed LS algorithm with centralized methods whereby, at each time instant k, all the n sensors transmit their raw data $\{y_{k+1,i}, \boldsymbol{\varphi}_{k,i}\}$ to a fusion center for processing to obtain a centralized estimate. Although the centralized algorithm may have some advantages over the distributed algorithm in terms of communication complexity, it also has some drawbacks compared with the distributed case. Firstly, the distributed methods may have stronger structural robustness compared with the centralized ones. This is because the centralized algorithm will fail once the fusion center is broken down by outside attacks or some sensors lost the connection to the fusion center, while the distributed algorithm can still estimate the unknown parameters even if the communications among some sensors are interrupted, as long as the network connectivity is maintained. Secondly, if the fusion center is far away from some sensors, the communications with the fusion center may not be feasible, and the transmission of observations and regression vectors may compromise the safety and privacy of the system even if the communication is possible. Hence, our distributed estimation problem is not a purely computational problem.

IV. Proofs of the main results

A. *Proof of* Theorem 3.1

To prove *Theorem 3.1*, we need to establish several lemmas first. The first lemma below is a key inequality on convex combination of nonnegative definite matrices.

Lemma 4.1: For any adjacency matrix $\mathcal{A} = \{a_{ij}\} \in \mathbb{R}^{n\times n}$, denote $\mathscr{A} = \mathcal{A}\otimes I_m$, and for any nonnegative definite matrices $Q_i \in \mathbb{R}^{m\times m}, i = 1,\ldots,n$, denote $Q = \text{diag}\{Q_1,\ldots,Q_n\}$, and $Q^{'} = \text{diag}\{Q_1^{'},\ldots,Q_n^{'}\}$, where $Q_i^{'} = \sum_{j=1}^{n} a_{ji}Q_j$. Then the following inequality holds:

$$\mathscr{A}Q\mathscr{A} \le Q^{'}. \tag{16}$$

Proof: By the definition of $\mathscr{A}$ and Q, we can get that

$$\mathscr{A}Q\mathscr{A} = \begin{pmatrix} \sum\limits_{j=1}^{n} a_{1j}a_{j1}Q_j & \cdots & \sum\limits_{j=1}^{n} a_{1j}a_{jn}Q_j \\ \vdots & \ddots & \vdots \\ \sum\limits_{j=1}^{n} a_{nj}a_{j1}Q_j & \cdots & \sum\limits_{j=1}^{n} a_{nj}a_{jn}Q_j \end{pmatrix}.$$

In order to prove the inequality, we only need to prove that for any unit column vector $x \in \mathbb{R}^{mn}$ with $\|x\| = 1$, $x^{\top}\mathscr{A}Q\mathscr{A}x \le x^{\top}Q^{'}x$ holds. Denote $x = \text{col}\{x_1, x_2, \ldots, x_n\}$ with $x_i \in \mathbb{R}^m$, then by the Schwarz inequality and noticing that $Q_j \ge 0$, $\sum_{j=1}^{n} a_{ij} = 1$, and $a_{ji} = a_{ij}, (i,j = 1,\ldots,n)$, we have

$$\begin{aligned}
x^{\top}\mathscr{A}Q\mathscr{A}x &= \sum_{p=1}^{n}\sum_{q=1}^{n}\sum_{j=1}^{n} a_{pj}a_{jq}x_p^{\top}Q_jx_q \\
&= \sum_{p=1}^{n}\sum_{q=1}^{n}\sum_{j=1}^{n} \sqrt{a_{pj}a_{jq}}x_p^{\top}Q_j^{\frac{1}{2}} \cdot \sqrt{a_{pj}a_{jq}}Q_j^{\frac{1}{2}}x_q \\
&\le \left\{\sum_{p=1}^{n}\sum_{q=1}^{n}\sum_{j=1}^{n} a_{pj}a_{jq}x_p^{\top}Q_jx_p\right\}^{\frac{1}{2}} \\
&\quad \cdot \left\{\sum_{p=1}^{n}\sum_{q=1}^{n}\sum_{j=1}^{n} a_{pj}a_{jq}x_q^{\top}Q_jx_q\right\}^{\frac{1}{2}} \\
&= \left\{\sum_{p=1}^{n}\sum_{j=1}^{n} a_{pj}x_p^{\top}Q_jx_p\right\}^{\frac{1}{2}} \left\{\sum_{q=1}^{n}\sum_{j=1}^{n} a_{jq}x_q^{\top}Q_jx_q\right\}^{\frac{1}{2}} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ji}x_i^{\top}Q_jx_i = x^{\top}Q^{'}x,
\end{aligned}$$

which completes the proof. ∎

By *Lemma 4.1*, we can obtain the following result:

Lemma 4.2: For any adjacency matrix $\mathcal{A} = \{a_{ij}\} \in \mathbb{R}^{n\times n}$, denote $\mathscr{A} = \mathcal{A}\otimes I_m$. Then for any $k \ge 1$,

$$\mathscr{A}\bar{\boldsymbol{P}}_{k+1}^{-1}\mathscr{A} \le \boldsymbol{P}_{k+1}^{-1}, \tag{17}$$

and

$$\mathscr{A}\boldsymbol{P}_{k+1}\mathscr{A} \le \bar{\boldsymbol{P}}_{k+1}, \tag{18}$$

holds, where $\bar{\boldsymbol{P}}_{k+1}$ and $\boldsymbol{P}_{k+1}$ are defined in (8).

Proof: By taking $Q_i = \bar{P}^{-1}_{k+1,i} \ge 0$ and noticing $P^{-1}_{k+1,i} = \sum_{j=1}^n a_{ji}\bar{P}^{-1}_{k+1,j} = Q_i^{'}$, we know from *Lemma 4.1* that

$$\mathscr{A}\bar{\boldsymbol{P}}^{-1}_{k+1}\mathscr{A} \le \boldsymbol{P}^{-1}_{k+1},$$

holds. To prove (18), we first assume that $\mathscr{A}$ is invertible. Then by *Lemma A.1* in Appendix A, it is easy to see that

$$\mathscr{A}\boldsymbol{P}_{k+1}\mathscr{A} \le \bar{\boldsymbol{P}}_{k+1}.$$

Next, we consider the case where $\mathscr{A}$ is not invertible. Since the number of eigenvalues of the matrix $\mathscr{A}$ is finite, then exists a constant $\varepsilon^* \in (0,1)$ such that the perturbed adjacency matrix $\mathscr{A}^\varepsilon = \mathscr{A} + \varepsilon I_{mn} = \{a^\varepsilon_{ij}\}$ will be invertible for any $0 < \varepsilon < \varepsilon^*$. By the definition of $\mathscr{A}^\varepsilon$, we know that $\mathscr{A}^\varepsilon$ is symmetric and the sums of each columns and rows of the matrix $\mathscr{A}^\varepsilon$ are all $1+\varepsilon$. Then we define $(P^\varepsilon_{k+1,i})^{-1} = \sum_{j=1}^n a^\varepsilon_{ji}\bar{P}^{-1}_{k+1,j}$, and we can denote $\boldsymbol{P}^\varepsilon_{k+1} = \mathrm{diag}\{P^\varepsilon_{k+1,1}, \dots, P^\varepsilon_{k+1,n}\}$ since $(P^\varepsilon_{k+1,i})^{-1}$ defined above is invertible. Similar to the proof of *Lemma 4.1*, for any unit column vector $x \in \mathbb{R}^{mn}$, we have

$$x^\top \mathscr{A}^\varepsilon \bar{\boldsymbol{P}}^{-1}_{k+1}\mathscr{A}^\varepsilon x \le (1+\varepsilon)x^\top(\boldsymbol{P}^\varepsilon_{k+1})^{-1}x.$$

Consequently, we have $\mathscr{A}^\varepsilon \bar{\boldsymbol{P}}^{-1}_{k+1}\mathscr{A}^\varepsilon \le (1+\varepsilon)(\boldsymbol{P}^\varepsilon_{k+1})^{-1}$. Since $\mathscr{A}^\varepsilon$ is invertible, we know from *Lemma A.1* in Appendix A that $\mathscr{A}^\varepsilon \boldsymbol{P}^\varepsilon_{k+1}\mathscr{A}^\varepsilon \le (1+\varepsilon)\bar{\boldsymbol{P}}_{k+1}$. By taking $\varepsilon \to 0$ on both sides of the above equation, we can obtain that

$$\lim_{\varepsilon\to 0}\mathscr{A}^\varepsilon \boldsymbol{P}^\varepsilon_{k+1}\mathscr{A}^\varepsilon = \mathscr{A}\boldsymbol{P}_{k+1}\mathscr{A} \le \lim_{\varepsilon\to 0}(1+\varepsilon)\bar{\boldsymbol{P}}_{k+1} = \bar{\boldsymbol{P}}_{k+1}.$$

This completes the proof. ∎

To accomplish the proof of *Theorem 3.1*, we also need the following inequality:

Lemma 4.3: For any adjacency matrix $\mathcal{A} = \{a_{ij}\} \in \mathbb{R}^{n\times n}$, and for any $k \ge 1$,

$$|\bar{\boldsymbol{P}}^{-1}_{k+1}| \le |\boldsymbol{P}^{-1}_{k+1}|, \tag{19}$$

holds, where $\bar{\boldsymbol{P}}_{k+1}$ and $\boldsymbol{P}_{k+1}$ are defined in (8).

Proof: By Ky Fan convex theorem [43] and noticing the definitions of $\bar{\boldsymbol{P}}_{k+1}, \boldsymbol{P}_{k+1}$, and $\mathcal{A} = \{a_{ij}\}$, we can see that

$$\begin{aligned}
|\boldsymbol{P}^{-1}_{k+1}| =& \prod_{i=1}^n \left|\sum_{j=1}^n a_{ji}\bar{P}^{-1}_{k+1,j}\right| \\
\ge& \prod_{i=1}^n |\bar{P}^{-1}_{k+1,1}|^{a_{1i}}|\bar{P}^{-1}_{k+1,2}|^{a_{2i}}\cdots|\bar{P}^{-1}_{k+1,n}|^{a_{ni}} \\
=& |\bar{P}^{-1}_{k+1,1}|^{\sum_{i=1}^n a_{1i}}|\bar{P}^{-1}_{k+1,2}|^{\sum_{i=1}^n a_{2i}}\cdots|\bar{P}^{-1}_{k+1,n}|^{\sum_{i=1}^n a_{ni}} \\
=& |\bar{P}^{-1}_{k+1,1}|\cdot|\bar{P}^{-1}_{k+1,2}|\cdots|\bar{P}^{-1}_{k+1,n}| = |\bar{\boldsymbol{P}}^{-1}_{k+1}|,
\end{aligned}$$

which completes the proof. ∎

To prove *Theorem 3.1*, we also need the following critical lemma:

Lemma 4.4: Let *Assumptions 3.1* and *3.2* be satisfied. Then the distributed LS defined by (7) and (8) satisfies the following relationship as $t \to \infty$:

$$\begin{aligned}
&\widetilde{\boldsymbol{\Theta}}^\top_{t+1}\boldsymbol{P}^{-1}_{t+1}\widetilde{\boldsymbol{\Theta}}_{t+1} + [1+o(1)]\sum_{k=0}^t \widetilde{\boldsymbol{\Theta}}^\top_k \boldsymbol{\Phi}_k \boldsymbol{b}_k \boldsymbol{\Phi}^\top_k \widetilde{\boldsymbol{\Theta}}_k \\
&\quad + [1+o(1)]\sum_{k=0}^t \widetilde{\boldsymbol{\Theta}}^\top_k \boldsymbol{P}^{-1}_k \boldsymbol{\Delta}_{k+1}\boldsymbol{P}^{-1}_k \widetilde{\boldsymbol{\Theta}}_k \\
&\le \sigma_w \log(|\boldsymbol{P}^{-1}_{t+1}|) + o(\log(|\boldsymbol{P}^{-1}_{t+1}|)) + O(1), \quad \text{a.s.},
\end{aligned} \tag{20}$$

where $\boldsymbol{\Delta}_{k+1} \triangleq \bar{\boldsymbol{P}}_{k+1} - \mathscr{A}\boldsymbol{P}_{k+1}\mathscr{A} \ge 0$ by *Lemma 4.2*, and $\sigma_w = \sum_{i=1}^n \sigma_i^2, \sigma_i^2 = \sup_{k\ge 0}\mathbb{E}[w^2_{k+1,i}|\mathcal{F}_k]$.

Proof: Since $\boldsymbol{b}_k = (I_n + \boldsymbol{\Phi}^\top_k \boldsymbol{P}_k \boldsymbol{\Phi}_k)^{-1}$ and $\boldsymbol{c}_k = \boldsymbol{b}_k \otimes I_m$, then by (9), we know that

$$\widetilde{\boldsymbol{\Theta}}_{k+1} = \boldsymbol{P}_{k+1}\mathscr{A}\boldsymbol{P}^{-1}_k\widetilde{\boldsymbol{\Theta}}_k - \boldsymbol{P}_{k+1}\mathscr{A}\bar{\boldsymbol{P}}^{-1}_{k+1}\boldsymbol{c}_k\boldsymbol{P}_k\boldsymbol{\Phi}_k\boldsymbol{W}_{k+1}.$$

Hence, we have the following expansion for the stochastic Lyapunov function $V_k = \widetilde{\boldsymbol{\Theta}}^\top_k \boldsymbol{P}^{-1}_k \widetilde{\boldsymbol{\Theta}}_k$:

$$\begin{aligned}
V_{k+1} =& \widetilde{\boldsymbol{\Theta}}^\top_{k+1}\boldsymbol{P}^{-1}_{k+1}\widetilde{\boldsymbol{\Theta}}_{k+1} \\
=& (\widetilde{\boldsymbol{\Theta}}^\top_k \boldsymbol{P}^{-1}_k \mathscr{A}\boldsymbol{P}_{k+1} - \boldsymbol{W}^\top_{k+1}\boldsymbol{\Phi}^\top_k\boldsymbol{P}_k\boldsymbol{c}_k\bar{\boldsymbol{P}}^{-1}_{k+1}\mathscr{A}\boldsymbol{P}_{k+1}) \\
&\cdot (\mathscr{A}\boldsymbol{P}^{-1}_k\widetilde{\boldsymbol{\Theta}}_k - \mathscr{A}\bar{\boldsymbol{P}}^{-1}_{k+1}\boldsymbol{c}_k\boldsymbol{P}_k\boldsymbol{\Phi}_k\boldsymbol{W}_{k+1}) \\
=& \widetilde{\boldsymbol{\Theta}}^\top_k\boldsymbol{P}^{-1}_k\mathscr{A}\boldsymbol{P}_{k+1}\mathscr{A}\boldsymbol{P}^{-1}_k\widetilde{\boldsymbol{\Theta}}_k \\
&- 2\widetilde{\boldsymbol{\Theta}}^\top_k\boldsymbol{P}^{-1}_k\mathscr{A}\boldsymbol{P}_{k+1}\mathscr{A}\bar{\boldsymbol{P}}^{-1}_{k+1}\boldsymbol{c}_k\boldsymbol{P}_k\boldsymbol{\Phi}_k\boldsymbol{W}_{k+1} \\
&+ \boldsymbol{W}^\top_{k+1}\boldsymbol{\Phi}^\top_k\boldsymbol{P}_k\boldsymbol{c}_k\bar{\boldsymbol{P}}^{-1}_{k+1}\mathscr{A}\boldsymbol{P}_{k+1}\mathscr{A}\bar{\boldsymbol{P}}^{-1}_{k+1} \\
&\cdot \boldsymbol{c}_k\boldsymbol{P}_k\boldsymbol{\Phi}_k\boldsymbol{W}_{k+1}.
\end{aligned} \tag{21}$$

Now, we proceed to estimate the right-hand-side (RHS) of (21) term by term. Firstly, we know that

$$\begin{aligned}
&\widetilde{\boldsymbol{\Theta}}^\top_k\boldsymbol{P}^{-1}_k\mathscr{A}\boldsymbol{P}_{k+1}\mathscr{A}\boldsymbol{P}^{-1}_k\widetilde{\boldsymbol{\Theta}}_k \\
=& \widetilde{\boldsymbol{\Theta}}^\top_k\boldsymbol{P}^{-1}_k\bar{\boldsymbol{P}}_{k+1}\boldsymbol{P}^{-1}_k\widetilde{\boldsymbol{\Theta}}_k - \widetilde{\boldsymbol{\Theta}}^\top_k\boldsymbol{P}^{-1}_k\boldsymbol{\Delta}_{k+1}\boldsymbol{P}^{-1}_k\widetilde{\boldsymbol{\Theta}}_k \\
=& \widetilde{\boldsymbol{\Theta}}^\top_k\boldsymbol{P}^{-1}_k(\boldsymbol{P}_k - \boldsymbol{P}_k\boldsymbol{\Phi}_k\boldsymbol{b}_k\boldsymbol{\Phi}^\top_k\boldsymbol{P}_k)\boldsymbol{P}^{-1}_k\widetilde{\boldsymbol{\Theta}}_k \\
&- \widetilde{\boldsymbol{\Theta}}^\top_k\boldsymbol{P}^{-1}_k\boldsymbol{\Delta}_{k+1}\boldsymbol{P}^{-1}_k\widetilde{\boldsymbol{\Theta}}_k \\
=& \widetilde{\boldsymbol{\Theta}}^\top_k\boldsymbol{P}^{-1}_k\widetilde{\boldsymbol{\Theta}}_k - \widetilde{\boldsymbol{\Theta}}^\top_k\boldsymbol{\Phi}_k\boldsymbol{b}_k\boldsymbol{\Phi}^\top_k\widetilde{\boldsymbol{\Theta}}_k - \widetilde{\boldsymbol{\Theta}}^\top_k\boldsymbol{P}^{-1}_k\boldsymbol{\Delta}_{k+1}\boldsymbol{P}^{-1}_k\widetilde{\boldsymbol{\Theta}}_k \\
=& V_k - \widetilde{\boldsymbol{\Theta}}^\top_k\boldsymbol{\Phi}_k\boldsymbol{b}_k\boldsymbol{\Phi}^\top_k\widetilde{\boldsymbol{\Theta}}_k - \widetilde{\boldsymbol{\Theta}}^\top_k\boldsymbol{P}^{-1}_k\boldsymbol{\Delta}_{k+1}\boldsymbol{P}^{-1}_k\widetilde{\boldsymbol{\Theta}}_k.
\end{aligned} \tag{22}$$

Moreover, by the (block) diagonal property of $\boldsymbol{b}_k, \boldsymbol{c}_k, \boldsymbol{P}_k$ and $\boldsymbol{\Phi}_k$, we have $\boldsymbol{c}_k\boldsymbol{P}_k = \boldsymbol{P}_k\boldsymbol{c}_k, \boldsymbol{\Phi}^\top_k\boldsymbol{c}_k = \boldsymbol{b}_k\boldsymbol{\Phi}^\top_k$, and $\boldsymbol{c}_k\boldsymbol{\Phi}_k = \boldsymbol{\Phi}_k\boldsymbol{b}_k$. By the matrix inversion lemma [36], we have

$$\bar{\boldsymbol{P}}^{-1}_{k+1} = \boldsymbol{P}^{-1}_k + \boldsymbol{\Phi}_k\boldsymbol{\Phi}^\top_k.$$

Thus, we can estimate the second term on the RHS of (21) as follows:

$$\begin{aligned}
&\widetilde{\boldsymbol{\Theta}}^\top_k\boldsymbol{P}^{-1}_k\mathscr{A}\boldsymbol{P}_{k+1}\mathscr{A}\bar{\boldsymbol{P}}^{-1}_{k+1}\boldsymbol{c}_k\boldsymbol{P}_k\boldsymbol{\Phi}_k\boldsymbol{W}_{k+1} \\
=& \widetilde{\boldsymbol{\Theta}}^\top_k\boldsymbol{P}^{-1}_k\mathscr{A}\boldsymbol{P}_{k+1}\mathscr{A}\boldsymbol{c}_k\boldsymbol{\Phi}_k\boldsymbol{W}_{k+1} \\
&+ \widetilde{\boldsymbol{\Theta}}^\top_k\boldsymbol{P}^{-1}_k\mathscr{A}\boldsymbol{P}_{k+1}\mathscr{A}\boldsymbol{\Phi}_k\boldsymbol{\Phi}^\top_k\boldsymbol{c}_k\boldsymbol{P}_k\boldsymbol{\Phi}_k\boldsymbol{W}_{k+1} \\
=& \widetilde{\boldsymbol{\Theta}}^\top_k\boldsymbol{P}^{-1}_k\mathscr{A}\boldsymbol{P}_{k+1}\mathscr{A}\boldsymbol{c}_k\boldsymbol{\Phi}_k\boldsymbol{W}_{k+1} \\
&+ \widetilde{\boldsymbol{\Theta}}^\top_k\boldsymbol{P}^{-1}_k\mathscr{A}\boldsymbol{P}_{k+1}\mathscr{A}\boldsymbol{\Phi}_k\boldsymbol{W}_{k+1} \\
&- \widetilde{\boldsymbol{\Theta}}^\top_k\boldsymbol{P}^{-1}_k\mathscr{A}\boldsymbol{P}_{k+1}\mathscr{A}\boldsymbol{\Phi}_k\boldsymbol{b}_k\boldsymbol{W}_{k+1} \\
=& \widetilde{\boldsymbol{\Theta}}^\top_k\boldsymbol{P}^{-1}_k\mathscr{A}\boldsymbol{P}_{k+1}\mathscr{A}\boldsymbol{\Phi}_k\boldsymbol{W}_{k+1} \\
=& \widetilde{\boldsymbol{\Theta}}^\top_k\boldsymbol{P}^{-1}_k\bar{\boldsymbol{P}}_{k+1}\boldsymbol{\Phi}_k\boldsymbol{W}_{k+1} - \widetilde{\boldsymbol{\Theta}}^\top_k\boldsymbol{P}^{-1}_k\boldsymbol{\Delta}_{k+1}\boldsymbol{\Phi}_k\boldsymbol{W}_{k+1}.
\end{aligned} \tag{23}$$

As for the last term on the RHS of (21), by $\mathscr{A}\boldsymbol{P}_{k+1}\mathscr{A} \leq \bar{\boldsymbol{P}}_{k+1}$, we can estimate it as follows:

$$\begin{aligned}
&\boldsymbol{W}_{k+1}^\top\boldsymbol{\Phi}_k^\top\boldsymbol{P}_k\boldsymbol{c}_k\bar{\boldsymbol{P}}_{k+1}^{-1}\mathscr{A}\boldsymbol{P}_{k+1}\mathscr{A}\bar{\boldsymbol{P}}_{k+1}^{-1}\boldsymbol{c}_k\boldsymbol{P}_k\boldsymbol{\Phi}_k\boldsymbol{W}_{k+1}\\
\leq&\boldsymbol{W}_{k+1}^\top\boldsymbol{\Phi}_k^\top\boldsymbol{P}_k\boldsymbol{c}_k(\boldsymbol{P}_k^{-1}+\boldsymbol{\Phi}_k\boldsymbol{\Phi}_k^\top)\boldsymbol{c}_k\boldsymbol{P}_k\boldsymbol{\Phi}_k\boldsymbol{W}_{k+1}\\
=&\boldsymbol{W}_{k+1}^\top\boldsymbol{\Phi}_k^\top\boldsymbol{P}_k\boldsymbol{c}_k^2\boldsymbol{\Phi}_k\boldsymbol{W}_{k+1}\\
&+\boldsymbol{W}_{k+1}^\top\boldsymbol{\Phi}_k^\top\boldsymbol{P}_k\boldsymbol{c}_k\boldsymbol{\Phi}_k\boldsymbol{\Phi}_k^\top\boldsymbol{c}_k\boldsymbol{P}_k\boldsymbol{\Phi}_k\boldsymbol{W}_{k+1}\\
=&\boldsymbol{W}_{k+1}^\top\boldsymbol{b}_k^2\boldsymbol{\Phi}_k^\top\boldsymbol{P}_k\boldsymbol{\Phi}_k\boldsymbol{W}_{k+1}\\
&+\boldsymbol{W}_{k+1}^\top(I_n+\boldsymbol{\Phi}_k^\top\boldsymbol{P}_k\boldsymbol{\Phi}_k)\boldsymbol{b}_k^2\boldsymbol{\Phi}_k^\top\boldsymbol{P}_k\boldsymbol{\Phi}_k\boldsymbol{W}_{k+1}\\
&-\boldsymbol{W}_{k+1}^\top\boldsymbol{b}_k^2\boldsymbol{\Phi}_k^\top\boldsymbol{P}_k\boldsymbol{\Phi}_k\boldsymbol{W}_{k+1}\\
=&\boldsymbol{W}_{k+1}^\top\boldsymbol{b}_k\boldsymbol{\Phi}_k^\top\boldsymbol{P}_k\boldsymbol{\Phi}_k\boldsymbol{W}_{k+1}.
\end{aligned}\tag{24}$$

By (21)-(24), we have

$$\begin{aligned}
V_{k+1}\leq&V_k-\widetilde{\boldsymbol{\Theta}}_k^\top\boldsymbol{\Phi}_k\boldsymbol{b}_k\boldsymbol{\Phi}_k^\top\widetilde{\boldsymbol{\Theta}}_k-\widetilde{\boldsymbol{\Theta}}_k^\top\boldsymbol{P}_k^{-1}\boldsymbol{\Delta}_{k+1}\boldsymbol{P}_k^{-1}\widetilde{\boldsymbol{\Theta}}_k\\
&-2\widetilde{\boldsymbol{\Theta}}_k^\top\boldsymbol{P}_k^{-1}\bar{\boldsymbol{P}}_{k+1}\boldsymbol{\Phi}_k\boldsymbol{W}_{k+1}\\
&+2\widetilde{\boldsymbol{\Theta}}_k^\top\boldsymbol{P}_k^{-1}\boldsymbol{\Delta}_{k+1}\boldsymbol{\Phi}_k\boldsymbol{W}_{k+1}\\
&+\boldsymbol{W}_{k+1}^\top\boldsymbol{b}_k\boldsymbol{\Phi}_k^\top\boldsymbol{P}_k\boldsymbol{\Phi}_k\boldsymbol{W}_{k+1}.
\end{aligned}\tag{25}$$

Summing from $k=0$ to t yields

$$\begin{aligned}
&V_{t+1}+\sum_{k=0}^{t}\widetilde{\boldsymbol{\Theta}}_k^\top\boldsymbol{\Phi}_k\boldsymbol{b}_k\boldsymbol{\Phi}_k^\top\widetilde{\boldsymbol{\Theta}}_k+\sum_{k=0}^{t}\widetilde{\boldsymbol{\Theta}}_k^\top\boldsymbol{P}_k^{-1}\boldsymbol{\Delta}_{k+1}\boldsymbol{P}_k^{-1}\widetilde{\boldsymbol{\Theta}}_k\\
\leq&V_0-2\sum_{k=0}^{t}\widetilde{\boldsymbol{\Theta}}_k^\top\boldsymbol{P}_k^{-1}\bar{\boldsymbol{P}}_{k+1}\boldsymbol{\Phi}_k\boldsymbol{W}_{k+1}\\
&-2\sum_{k=0}^{t}\widetilde{\boldsymbol{\Theta}}_k^\top\boldsymbol{P}_k^{-1}(-\boldsymbol{\Delta}_{k+1})\boldsymbol{\Phi}_k\boldsymbol{W}_{k+1}\\
&+\sum_{k=0}^{t}\boldsymbol{W}_{k+1}^\top\boldsymbol{b}_k\boldsymbol{\Phi}_k^\top\boldsymbol{P}_k\boldsymbol{\Phi}_k\boldsymbol{W}_{k+1}.
\end{aligned}\tag{26}$$

Next, we estimate the last three terms on the RHS of (26) separately. By *Assumptions 3.1* and *3.2*, and $\widetilde{\boldsymbol{\Theta}}_k^\top\boldsymbol{P}_k^{-1}\bar{\boldsymbol{P}}_{k+1}\boldsymbol{\Phi}_k\in\mathcal{F}_k$, $\widetilde{\boldsymbol{\Theta}}_k^\top\boldsymbol{P}_k^{-1}(-\boldsymbol{\Delta}_{k+1})\boldsymbol{\Phi}_k\in\mathcal{F}_k$, we can use the martingale estimation theorem (*Theorem 2.8* in [28]) to get the following estimation for any $\delta>0$:

$$\begin{aligned}
&\sum_{k=0}^{t}\widetilde{\boldsymbol{\Theta}}_k^\top\boldsymbol{P}_k^{-1}\bar{\boldsymbol{P}}_{k+1}\boldsymbol{\Phi}_k\boldsymbol{W}_{k+1}\\
=&O\bigg(\Big\{\sum_{k=0}^{t}\|\widetilde{\boldsymbol{\Theta}}_k^\top\boldsymbol{P}_k^{-1}\bar{\boldsymbol{P}}_{k+1}\boldsymbol{\Phi}_k\|^2\Big\}^{\frac{1}{2}+\delta}\bigg),\quad\text{a.s.,}
\end{aligned}\tag{27}$$

and

$$\begin{aligned}
&\sum_{k=0}^{t}\widetilde{\boldsymbol{\Theta}}_k^\top\boldsymbol{P}_k^{-1}(-\boldsymbol{\Delta}_{k+1})\boldsymbol{\Phi}_k\boldsymbol{W}_{k+1}\\
=&O\bigg(\Big\{\sum_{k=0}^{t}\|\widetilde{\boldsymbol{\Theta}}_k^\top\boldsymbol{P}_k^{-1}\boldsymbol{\Delta}_{k+1}\boldsymbol{\Phi}_k\|^2\Big\}^{\frac{1}{2}+\delta}\bigg),\quad\text{a.s.}
\end{aligned}\tag{28}$$

To further analyze (27) and (28), we note from the definitions of $\bar{\boldsymbol{P}}_{k+1}$ and $\boldsymbol{b}_k$ that

$$\begin{aligned}
&\boldsymbol{P}_k^{-1}\bar{\boldsymbol{P}}_{k+1}\boldsymbol{\Phi}_k=\boldsymbol{\Phi}_k-\boldsymbol{c}_k\boldsymbol{\Phi}_k\boldsymbol{\Phi}_k^\top\boldsymbol{P}_k\boldsymbol{\Phi}_k\\
=&\boldsymbol{\Phi}_k-\boldsymbol{c}_k\boldsymbol{\Phi}_k(I_n+\boldsymbol{\Phi}_k^\top\boldsymbol{P}_k\boldsymbol{\Phi}_k)+\boldsymbol{c}_k\boldsymbol{\Phi}_k=\boldsymbol{\Phi}_k\boldsymbol{b}_k.
\end{aligned}$$

Hence, it is easy to see that

$$\begin{aligned}
\|\widetilde{\boldsymbol{\Theta}}_k^\top\boldsymbol{P}_k^{-1}\bar{\boldsymbol{P}}_{k+1}\boldsymbol{\Phi}_k\|^2=&\widetilde{\boldsymbol{\Theta}}_k^\top\boldsymbol{P}_k^{-1}\bar{\boldsymbol{P}}_{k+1}\boldsymbol{\Phi}_k\boldsymbol{\Phi}_k^\top\bar{\boldsymbol{P}}_{k+1}\boldsymbol{P}_k^{-1}\widetilde{\boldsymbol{\Theta}}_k\\
=&\widetilde{\boldsymbol{\Theta}}_k^\top\boldsymbol{\Phi}_k\boldsymbol{b}_k^2\boldsymbol{\Phi}_k^\top\widetilde{\boldsymbol{\Theta}}_k\leq\widetilde{\boldsymbol{\Theta}}_k^\top\boldsymbol{\Phi}_k\boldsymbol{b}_k\boldsymbol{\Phi}_k^\top\widetilde{\boldsymbol{\Theta}}_k,
\end{aligned}$$

By taking $0<\delta<\frac{1}{2}$, we have from (27) that

$$\begin{aligned}
&\sum_{k=0}^{t}\widetilde{\boldsymbol{\Theta}}_k^\top\boldsymbol{P}_k^{-1}\bar{\boldsymbol{P}}_{k+1}\boldsymbol{\Phi}_k\boldsymbol{W}_{k+1}\\
=&O(1)+o\bigg(\Big\{\sum_{k=0}^{t}\|\widetilde{\boldsymbol{\Theta}}_k^\top\boldsymbol{P}_k^{-1}\bar{\boldsymbol{P}}_{k+1}\boldsymbol{\Phi}_k\|^2\Big\}\bigg)\\
=&O(1)+o\bigg(\Big\{\sum_{k=0}^{t}\widetilde{\boldsymbol{\Theta}}_k^\top\boldsymbol{\Phi}_k\boldsymbol{b}_k\boldsymbol{\Phi}_k^\top\widetilde{\boldsymbol{\Theta}}_k\Big\}\bigg),\quad\text{a.s.}
\end{aligned}\tag{29}$$

Moreover, since $\boldsymbol{\Delta}_{k+1}=\bar{\boldsymbol{P}}_{k+1}-\mathscr{A}\boldsymbol{P}_{k+1}\mathscr{A}\leq\bar{\boldsymbol{P}}_{k+1}$, then

$$\begin{aligned}
&\boldsymbol{\Delta}_{k+1}^{\frac{1}{2}}\boldsymbol{\Phi}_k\boldsymbol{\Phi}_k^\top\boldsymbol{\Delta}_{k+1}^{\frac{1}{2}}\leq\lambda_{\max}\{\boldsymbol{\Phi}_k^\top\boldsymbol{\Delta}_{k+1}\boldsymbol{\Phi}_k\}\cdot I_{mn}\\
\leq&\lambda_{\max}\{\boldsymbol{\Phi}_k^\top\bar{\boldsymbol{P}}_{k+1}\boldsymbol{\Phi}_k\}\cdot I_{mn}\\
=&\lambda_{\max}\{\boldsymbol{\Phi}_k^\top(\boldsymbol{P}_k-\boldsymbol{c}_k\boldsymbol{P}_k\boldsymbol{\Phi}_k\boldsymbol{\Phi}_k^\top\boldsymbol{P}_k)\boldsymbol{\Phi}_k\}\cdot I_{mn}\\
=&\lambda_{\max}\{\boldsymbol{b}_k\boldsymbol{\Phi}_k^\top\boldsymbol{P}_k\boldsymbol{\Phi}_k\}\cdot I_{mn}<I_{mn}.
\end{aligned}$$

Hence, we have $\boldsymbol{\Delta}_{k+1}\boldsymbol{\Phi}_k\boldsymbol{\Phi}_k^\top\boldsymbol{\Delta}_{k+1}\leq\boldsymbol{\Delta}_{k+1}$, and so we have

$$\begin{aligned}
\|\widetilde{\boldsymbol{\Theta}}_k^\top\boldsymbol{P}_k^{-1}\boldsymbol{\Delta}_{k+1}\boldsymbol{\Phi}_k\|^2=&\widetilde{\boldsymbol{\Theta}}_k^\top\boldsymbol{P}_k^{-1}\boldsymbol{\Delta}_{k+1}\boldsymbol{\Phi}_k\boldsymbol{\Phi}_k^\top\boldsymbol{\Delta}_{k+1}\boldsymbol{P}_k^{-1}\widetilde{\boldsymbol{\Theta}}_k\\
\leq&\widetilde{\boldsymbol{\Theta}}_k^\top\boldsymbol{P}_k^{-1}\boldsymbol{\Delta}_{k+1}\boldsymbol{P}_k^{-1}\widetilde{\boldsymbol{\Theta}}_k.
\end{aligned}\tag{30}$$

By taking $0<\delta<\frac{1}{2}$, we know from (28) that

$$\begin{aligned}
&\sum_{k=0}^{t}\widetilde{\boldsymbol{\Theta}}_k^\top\boldsymbol{P}_k^{-1}(-\boldsymbol{\Delta}_{k+1})\boldsymbol{\Phi}_k\boldsymbol{W}_{k+1}\\
=&O(1)+o\bigg(\Big\{\sum_{k=0}^{t}\|\widetilde{\boldsymbol{\Theta}}_k^\top\boldsymbol{P}_k^{-1}\boldsymbol{\Delta}_{k+1}\boldsymbol{\Phi}_k\|^2\Big\}\bigg)\\
=&O(1)+o\bigg(\Big\{\sum_{k=0}^{t}\widetilde{\boldsymbol{\Theta}}_k^\top\boldsymbol{P}_k^{-1}\boldsymbol{\Delta}_{k+1}\boldsymbol{P}_k^{-1}\widetilde{\boldsymbol{\Theta}}_k\Big\}\bigg),\quad\text{a.s.}
\end{aligned}\tag{31}$$

We now proceed to estimate the last term in (26). Firstly, we know that

$$\begin{aligned}
&\boldsymbol{W}_{k+1}^\top\boldsymbol{b}_k\boldsymbol{\Phi}_k^\top\boldsymbol{P}_k\boldsymbol{\Phi}_k\boldsymbol{W}_{k+1}\leq\|\boldsymbol{b}_k\boldsymbol{\Phi}_k^\top\boldsymbol{P}_k\boldsymbol{\Phi}_k\|\cdot\|\boldsymbol{W}_{k+1}\|^2\\
=&\lambda_{\max}\{\boldsymbol{b}_k\boldsymbol{\Phi}_k^\top\boldsymbol{P}_k\boldsymbol{\Phi}_k\}\cdot\Big\{\sum_{i=1}^{n}w_{k+1,i}^2\Big\}.
\end{aligned}\tag{32}$$

Following a similar proof idea as in the traditional single sensor case ([25], see also [28]), from $\bar{\boldsymbol{P}}_{k+1}=\boldsymbol{P}_k-\boldsymbol{c}_k\boldsymbol{P}_k\boldsymbol{\Phi}_k\boldsymbol{\Phi}_k^\top\boldsymbol{P}_k$, we have $\boldsymbol{P}_k^{-1}=\bar{\boldsymbol{P}}_{k+1}^{-1}(I_{mn}-\boldsymbol{c}_k\boldsymbol{P}_k\boldsymbol{\Phi}_k\boldsymbol{\Phi}_k^\top)$. By taking determinants on both sides of the above identity, and noticing $0\leq\boldsymbol{b}_k\boldsymbol{\Phi}_k^\top\boldsymbol{P}_k\boldsymbol{\Phi}_k\leq I_n$, we have

$$\begin{aligned}
|\boldsymbol{P}_k^{-1}|=&|\bar{\boldsymbol{P}}_{k+1}^{-1}|\cdot|I_n-\boldsymbol{b}_k\boldsymbol{\Phi}_k^\top\boldsymbol{P}_k\boldsymbol{\Phi}_k|\\
\leq&|\bar{\boldsymbol{P}}_{k+1}^{-1}|\cdot(1-\lambda_{\max}\{\boldsymbol{b}_k\boldsymbol{\Phi}_k^\top\boldsymbol{P}_k\boldsymbol{\Phi}_k\}).
\end{aligned}$$

Moreover, we know from *Lemma 4.3* that

$$\lambda_{\max}\{\boldsymbol{b}_k\boldsymbol{\Phi}_k^\top\boldsymbol{P}_k\boldsymbol{\Phi}_k\}\leq\frac{|\bar{\boldsymbol{P}}_{k+1}^{-1}|-|\boldsymbol{P}_k^{-1}|}{|\bar{\boldsymbol{P}}_{k+1}^{-1}|}\leq\frac{|\boldsymbol{P}_{k+1}^{-1}|-|\boldsymbol{P}_k^{-1}|}{|\boldsymbol{P}_{k+1}^{-1}|}.$$

Therefore

$$\sum_{k=0}^{t}\lambda_{\max}\{\boldsymbol{b}_k\boldsymbol{\Phi}_k^\top\boldsymbol{P}_k\boldsymbol{\Phi}_k\}\le\sum_{k=0}^{t}\frac{|\boldsymbol{P}_{k+1}^{-1}|-|\boldsymbol{P}_k^{-1}|}{|\boldsymbol{P}_{k+1}^{-1}|}$$
$$\le\sum_{k=0}^{t}\int_{|\boldsymbol{P}_k^{-1}|}^{|\boldsymbol{P}_{k+1}^{-1}|}\frac{dx}{x}=\log(|\boldsymbol{P}_{t+1}^{-1}|)-\log(|\boldsymbol{P}_0^{-1}|).\qquad(33)$$

Consequently, by using the martingale estimation theorem (*Theorem 2.8* in [28]), we have for any $\forall\eta>0$,

$$\sum_{k=0}^{t}\lambda_{\max}\{\boldsymbol{b}_k\boldsymbol{\Phi}_k^\top\boldsymbol{P}_k\boldsymbol{\Phi}_k\}\Big\{\sum_{i=1}^{n}w_{k+1,i}^2-\mathbb{E}\Big[\sum_{i=1}^{n}w_{k+1,i}^2|\mathcal{F}_k\Big]\Big\}$$
$$=O\bigg(S_t\Big(\frac{\beta}{2}\Big)\Big\{\log\Big(S_t\Big(\frac{\beta}{2}\Big)+e\Big)\Big\}^{\frac{2}{\beta}+\eta}\bigg),\quad\text{a.s.},\qquad(34)$$

where

$$S_t\Big(\frac{\beta}{2}\Big)\triangleq\Big[\sum_{k=0}^{t}(\lambda_{\max}\{\boldsymbol{b}_k\boldsymbol{\Phi}_k^\top\boldsymbol{P}_k\boldsymbol{\Phi}_k\})^{\frac{\beta}{2}}\Big]^{\frac{2}{\beta}}.$$

Since $\boldsymbol{b}_k\boldsymbol{\Phi}_k^\top\boldsymbol{P}_k\boldsymbol{\Phi}_k\le I_n$ and $\frac{\beta}{2}>1$, we have from (33) that

$$S_t\Big(\frac{\beta}{2}\Big)=O(1)+O((\log|\boldsymbol{P}_{t+1}^{-1}|)^{\frac{2}{\beta}}).$$

From this, we can get from (32)-(34) that

$$\sum_{k=0}^{t}\boldsymbol{W}_{k+1}^\top\boldsymbol{b}_k\boldsymbol{\Phi}_k^\top\boldsymbol{P}_k\boldsymbol{\Phi}_k\boldsymbol{W}_{k+1}$$
$$\le\sum_{i=1}^{n}\sigma_i^2\sum_{k=0}^{t}\lambda_{\max}\{\boldsymbol{b}_k\boldsymbol{\Phi}_k^\top\boldsymbol{P}_k\boldsymbol{\Phi}_k\}+o(\log|\boldsymbol{P}_{t+1}^{-1}|)+O(1)$$
$$\le\sigma_w\log|\boldsymbol{P}_{t+1}^{-1}|+o(\log|\boldsymbol{P}_{t+1}^{-1}|)+O(1).$$

Finally, substituting this into (26), we know that the desired result (20) is true. This completes the proof. ■

Proof of Theorem 3.1:

By the definitions of $\bar{P}_{t,i}^{-1}$ and $P_{t,i}^{-1}$, it is easy to know that for any $t\ge0$,

$$P_{t+1,i}^{-1}=\sum_{j=1}^{n}a_{ji}\bar{P}_{t+1,j}^{-1}=\sum_{j=1}^{n}a_{ji}(P_{t,j}^{-1}+\boldsymbol{\varphi}_{t,j}\boldsymbol{\varphi}_{t,j}^\top).$$

Consequently, we have

$$\max_{1\le i\le n}\lambda_{\max}\{P_{t+1,i}^{-1}\}$$
$$=\max_{i=1,\dots,n}\lambda_{\max}\Big\{\sum_{j=1}^{n}a_{ji}(P_{t,j}^{-1}+\boldsymbol{\varphi}_{t,j}\boldsymbol{\varphi}_{t,j}^\top)\Big\}$$
$$\le\max_{1\le i\le n}\sum_{j=1}^{n}a_{ji}\Big(\lambda_{\max}\{P_{t,j}^{-1}\}+\lambda_{\max}\{\boldsymbol{\varphi}_{t,j}\boldsymbol{\varphi}_{t,j}^\top\}\Big)$$
$$\le\max_{1\le i\le n}\lambda_{\max}\{P_{t,i}^{-1}\}+\sum_{j=1}^{n}\|\boldsymbol{\varphi}_{t,j}\|^2$$
$$\le\max_{1\le i\le n}\lambda_{\max}\{P_{0,i}^{-1}\}+\sum_{j=1}^{n}\sum_{k=0}^{t}\|\boldsymbol{\varphi}_{k,j}\|^2$$
$$\le\lambda_{\max}\{\boldsymbol{P}_0^{-1}\}+\sum_{j=1}^{n}\sum_{k=0}^{t}\|\boldsymbol{\varphi}_{k,j}\|^2.\qquad(35)$$

From (35) and the connection between determinant and eigenvalues of a matrix, it is easy to conclude that

$$\log(|\boldsymbol{P}_{t+1}^{-1}|)\le mn\log\Big(\max_{1\le i\le n}\lambda_{\max}\{P_{t+1,i}^{-1}\}\Big)\le mn\log(r_t).$$

Consequently, *Theorem 3.1* follows from this and *Lemma 4.4* immediately.

B. Proof of Theorem 3.2

By the definition of $\boldsymbol{b}_k$ in (8), we know that

$$\boldsymbol{\Phi}_k\boldsymbol{\Phi}_k^\top=\boldsymbol{\Phi}_k\boldsymbol{b}_k\boldsymbol{\Phi}_k^\top+\boldsymbol{\Phi}_k(\boldsymbol{b}_k\boldsymbol{\Phi}_k^\top\boldsymbol{P}_k\boldsymbol{\Phi}_k)\boldsymbol{\Phi}_k^\top.$$

Then by noticing that $\boldsymbol{b}_k,\boldsymbol{\Phi}_k$ and $\boldsymbol{P}_k$ are (block) diagonal matrices, and $\boldsymbol{\Phi}_k^\top\boldsymbol{P}_k\boldsymbol{\Phi}_k=O(1)$, a.s., we know that

$$\sum_{i=1}^{n}\sum_{k=0}^{t}R_{k,i}=\sum_{i=1}^{n}\sum_{k=0}^{t}(\boldsymbol{\varphi}_{k,i}^\top\widetilde{\boldsymbol{\theta}}_{k,i})^2=\sum_{k=0}^{t}\widetilde{\boldsymbol{\Theta}}_k^\top\boldsymbol{\Phi}_k\boldsymbol{\Phi}_k^\top\widetilde{\boldsymbol{\Theta}}_k$$
$$=\sum_{k=0}^{t}\widetilde{\boldsymbol{\Theta}}_k^\top\boldsymbol{\Phi}_k\boldsymbol{b}_k\boldsymbol{\Phi}_k^\top\widetilde{\boldsymbol{\Theta}}_k+\sum_{k=0}^{t}\widetilde{\boldsymbol{\Theta}}_k^\top\boldsymbol{\Phi}_k(\boldsymbol{b}_k\boldsymbol{\Phi}_k^\top\boldsymbol{P}_k\boldsymbol{\Phi}_k)\boldsymbol{\Phi}_k^\top\widetilde{\boldsymbol{\Theta}}_k$$
$$=O\bigg(\sum_{k=0}^{t}\widetilde{\boldsymbol{\Theta}}_k^\top\boldsymbol{\Phi}_k\boldsymbol{b}_k\boldsymbol{\Phi}_k^\top\widetilde{\boldsymbol{\Theta}}_k\bigg).\qquad(36)$$

Substituting this into *Theorem 3.1* 1), we conclude that (12) holds.

C. Proof of Theorem 3.3

For ease of representation, let $a_{ij}^{(s)}$ be the (i,j)-th entry of the matrix $\mathcal{A}^s, s\ge1$. Note that $a_{ij}^{(1)}=a_{ij}$. By *Assumption 3.3* and *Remark 3.1*, we know that $a_{ji}^{(D_\mathcal{G})}\ge a_{\min}>0$, where $a_{\min}=\min_{i,j\in\mathcal{V}}a_{ij}^{(D_\mathcal{G})}>0$, and $D_\mathcal{G}$ is the diameter of the graph $\mathcal{G}$. Consequently, it is not difficult to see that for any $k>D_\mathcal{G}$, $a_{ji}^{(k)}\ge a_{\min}$ holds.

By (8), it is easy to see that for any $t\ge0$,

$$\text{vec}\{\boldsymbol{P}_{t+1}^{-1}\}=\mathscr{A}\text{vec}\{\bar{\boldsymbol{P}}_{t+1}^{-1}\}=\mathscr{A}\text{vec}\{\boldsymbol{P}_t^{-1}\}+\mathscr{A}\text{vec}\{\boldsymbol{\Phi}_t\boldsymbol{\Phi}_t^\top\}$$
$$=\mathscr{A}^{t+1}\text{vec}\{\boldsymbol{P}_0^{-1}\}+\sum_{k=0}^{t}\mathscr{A}^{t-k+1}\text{vec}\{\boldsymbol{\Phi}_k\boldsymbol{\Phi}_k^\top\},$$

which implies that for any $t\ge D_\mathcal{G}$,

$$P_{t+1,i}^{-1}=\sum_{j=1}^{n}a_{ji}^{(t+1)}P_{0,j}^{-1}+\sum_{j=1}^{n}\sum_{k=0}^{t}a_{ji}^{(t-k+1)}\boldsymbol{\varphi}_{k,j}\boldsymbol{\varphi}_{k,j}^\top$$
$$\ge\sum_{j=1}^{n}a_{ji}^{(t+1)}P_{0,j}^{-1}+\sum_{j=1}^{n}\sum_{k=0}^{t-D_\mathcal{G}+1}a_{ji}^{(t-k+1)}\boldsymbol{\varphi}_{k,j}\boldsymbol{\varphi}_{k,j}^\top$$
$$\ge a_{\min}\sum_{j=1}^{n}P_{0,j}^{-1}+a_{\min}\sum_{j=1}^{n}\sum_{k=0}^{t-D_\mathcal{G}+1}\boldsymbol{\varphi}_{k,j}\boldsymbol{\varphi}_{k,j}^\top.\qquad(37)$$

Then we have

$$\lambda_{\min}\{\boldsymbol{P}_{t+1}^{-1}\}\ge a_{\min}\lambda_{\min}\bigg\{\sum_{j=1}^{n}P_{0,j}^{-1}+\sum_{j=1}^{n}\sum_{k=0}^{t-D_\mathcal{G}+1}\boldsymbol{\varphi}_{k,j}\boldsymbol{\varphi}_{k,j}^\top\bigg\}.$$

Note also that

$$\|\widetilde{\boldsymbol{\Theta}}_{t+1}\|^2\le\widetilde{\boldsymbol{\Theta}}_{t+1}^\top\bigg[\frac{\boldsymbol{P}_{t+1}^{-1}}{\lambda_{\min}\{\boldsymbol{P}_{t+1}^{-1}\}}\bigg]\widetilde{\boldsymbol{\Theta}}_{t+1}.$$

Hence, by 2) in *Theorem 3.1*, we know that *Theorem 3.3* holds.

V. Concluding remarks

In this paper, we have established a convergence theory for a basic class of distributed LS algorithms, under quite general conditions on the measured information or data used in the estimation. The accumulated regret of adaptive predictors has been shown to have a celebrated logarithm increase without any excitation condition imposed on the system data, and the convergence rate of the distributed LS estimates has also been established under a cooperative excitation condition, which can be regarded as an extension of the weakest possible excitation condition known for the convergence of the classical LS. Neither independence and stationarity, nor Gaussian property, are required in our results. Moreover, the cooperative excitation condition introduced and used in the paper indicates that the distributed LS can fulfill the estimation task cooperatively, even if any individual sensor cannot due to lack of necessary excitation.

References

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

[2] A. H. Sayed, "Adaptive networks," *Proceedings of the IEEE*, vol. 102, no. 4, pp. 460–497, April 2014.

[3] S. Y. Xie and L. Guo, "A necessary and sufficient condition for stability of LMS-based consensus adaptive filters," *Automatica*, vol. 93, pp. 12–19, July 2018.

[4] S. Y. Xie and L. Guo, "Analysis of normalized least mean squares-based consensus adaptive filters under a general information condition," *SIAM Journal on Control and Optimization*, vol. 56, no. 5, pp. 3404–3431, Sept. 2018.

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

[6] I. E. K. Harrane, R. Flamary and C. Richard, "On reducing the communication cost of the diffusion LMS algorithm," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 5, no. 1, pp. 100–112, March 2019.

[7] A. H. Sayed and C. G. Lopes, "Distributed recursive least-squares strategies over adaptive networks," in *40th Asilomar Conference on Signals, Systems and Computers*, Pacific, Grove, CA, Oct. 2006, pp. 233–237.

[8] G. Mateos and G. B. Giannakis, "Distributed recursive least-squares: stability and performance analysis," *IEEE Transactions on Signal Processing*, vol. 60, no. 7, pp. 3740–3754, July 2012.

[9] L. Xiao, S. Boyd, and S. Lall, "A space-time diffusion scheme for peer-to-peer least-squares estimation," in *Proceedings of 5th International Conference on Information Processing in Sensor Networks (IPSN 2006)*, Nashville, TN, April 2006, pp. 168–176.

[10] F. S. Cattivelli, C. G. Lopes, and A. H. Sayed, "Diffusion recursive least squares for distributed estimation over adaptive networks," *IEEE Transactions on Signal Processing*, vol. 56, no. 5, pp. 1865–1877, May 2008.

[11] A. Bertrand, M. Moonen, and A. H. Sayed, "Diffusion bias-compensated RLS estimation over adaptive networks," *IEEE Transactions on Signal Processing*, vol. 59, no. 11, pp. 5212–5224, Nov. 2011.

[12] R. Arablouei, K. Dogancay, S. Werner, and Y-F Huang, "Adaptive distributed estimation based on recursive least-squares and partial diffusion," *IEEE Transactions on Signal Processing*, vol. 62, no. 14, pp. 3510–3522, July 2014.

[13] V. Vahidpour, A. Rastegarnia, A. Khalili and S. Sanei, "Analysis of partial diffusion recursive least squares adaptation over noisy links," *IET Signal Processing*, vol. 11, no. 6, pp. 749–757, August 2017.

[14] A. Rastegarnia, "Reduced-communication diffusion RLS for distributed estimation over multi-agent networks," *IEEE Transactions on Circuits and Systems II: Express Briefs*, doi: 10.1109/TCSII.2019.2899194.

[15] Y. Yu, H. Zhao, R. C. de Lamare, Y. Zakharov and L. Lu, "Robust distributed diffusion recursive least squares algorithms with side information for adaptive networks," *IEEE Transactions on Signal Processing*, vol. 67, no. 6, pp. 1566–1581, March 2019.

[16] R. Carli, A. Chiuso, L. Schenato, and S. Zampieri, "Distributed Kalman filtering based on consensus strategies," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 4, pp. 622–633, May 2008.

[17] G. Battistelli, L. Chisci, G. Mugnai, A. Farina, and A. Graziano, "Consensus-based linear and nonlinear filtering," *IEEE Transactions on Automatic Control*, vol. 60, no. 5, pp. 1410–1415, May 2015.

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

[19] S. Das and J. Moura, "Consensus+innovations distributed Kalman filter with optimized gains," *IEEE Transactions on Signal Processing*, vol. 65, no. 2, pp. 467–481, Jan. 2017.

[20] K. J. Åström and B. Wittenmark, "On self tuning regulators," *Automatica*, vol. 9, no. 2, pp. 185–199, March 1973.

[21] L. Guo, "Convergence and logarithm laws of self-tuning regulators," *Automatica*, vol. 31, no. 3, pp. 435–450, March 1995.

[22] L. Ljung, "Consistency of the least-squares identification method," *IEEE Transactions on Automatic Control*, vol. 21, no. 5, pp. 779–781, Oct. 1976.

[23] J. B. Moore, "On strong consistency of least squares identification algorithm," *Automatica*, vol. 14, no. 5, pp. 505–509, Sep. 1978.

[24] H. F. Chen, "Strong consistency and convergence rate of least squares identification," *Science in China Series A - Mathematics, Physics, Astronomy & Technological Science*, vol. 25, no. 7, pp. 771–784, 1982.

[25] T. L. Lai and C. Z. Wei, "Least squares estimates in stochastic regression models with applications to identification and control dynamic systems," *Annals of Statistics*, vol. 10, no. 1, pp. 154–166, March 1982.

[26] T. L. Lai and C. Z. Wei, "Extended least squares and their applications to adaptive control and prediction in linear systems," *IEEE Transactions on Automatic Control*, vol. 31, no. 10, pp. 898–906, Oct. 1986.

[27] H. F. Chen and L. Guo, "Convergence rate of least-squares identification and adaptive control for stochastic systems," *International Journal of Control*, vol. 44, no. 5, pp. 1459–1476, Nov. 1986.

[28] H. F. Chen and L. Guo, *Identification and Stochastic Adaptive Control*, Birkhäsuser, Boston, 1991.

[29] L. Guo and H. F. Chen, "The Åström-Wittenmark self-tuning regulator revised and ELS-based adaptive tracker," *IEEE Transactions on Automatic Control*, vol. 36, no. 7, pp. 802–812, July 1991.

[30] L. Ljung, "Analysis of recursive stochastic algorithms," *IEEE Transactions on Automatic Control*, vol. 22, no. 4, pp. 551–575, Aug. 1977.

[31] G. C. Goodwin, P. J. Ramadge, and P. E. Caines, "Discrete time stochastic adaptive control," *SIAM Journal on Control and Optimization*, vol. 19, no. 6, pp. 829–853, 1981.

[32] P. R. Kumar, "Convergence of adaptive control schemes using least-squares parameter estimates," *IEEE Transactions on Automatic Control*, vol. 35, no. 4, pp. 416–424, April 1990.

[33] S. Julier and J. Uhlmann, "A non-divergent estimation algorithm in the presence of unknown correlations," in *Proceedings of the 1997 American Control Conference*, Albuquerque, NM, USA, April 1997, pp. 2369–2373.

[34] L. Chen, P. Arambel, and P. Mehra, "Estimation under unknown correlation: covariance intersection revisited," *IEEE Transactions on Automatic Control*, vol. 47, no. 11, pp. 1879–1882, May 2002.

[35] Y. S. Chow and H. Teicher, *Probability Theory*, New York: Springer, March 1978.

[36] L. Guo, *Time-Varying Stochastic Systems–Stability and Adaptive Theory (Second Edition)*, Science Press, 2020.

[37] M. Zhong and C. G. Cassandras, "Asynchronous distributed optimization with event-driven communication," *IEEE Transactions on Automatic Control*, vol. 55, no. 12, pp. 2735–2750, Dec. 2010.

[38] S. Y. Xie and L. Guo, "Analysis of compressed distributed adaptive filters," *Automatica*, vol. 112, 108707, Feb. 2020.

[39] C. Godsil and G. Royle, *Algebraic Graph Theory*, Springer-Verlag, 2014.

[40] S. Hosseini, A. Chapman, and M. Mesbahi, "Online distributed convex optimization on dynamic networks," *IEEE Transactions on Automatic Control*, vol. 61, no. 11, pp. 3545–3550, Nov. 2016.

[41] S. Shahrampour and A. Jadbabaie, "Distributed online optimization in dynamic environments using mirror descent," *IEEE Transactions on Automatic Control*, vol. 63, no. 3, pp. 714–725, March 2018.

[42] T. L. Lai, "Asymptotically efficient adaptive control in stochastic regression models," *Advances in Applied Mathematics*, vol. 7, no. 1, pp. 23–45, March 1986.

[43] K. Fan, "On a theorem of Weyl concerning eigenvalues of linear transformations," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 36, no. 1, pp. 31–35, Jan. 1950.