Preprints of the 19th World Congress
The International Federation of Automatic Control
Cape Town, South Africa. August 24-29, 2014

Randomized Incremental Least Squares for Distributed Estimation Over Sensor Networks ⋆

Keyou You *, **Shiji Song** * and **Li Qiu** **

* *Department of Automation, Tsinghua University, Beijing, 100084, China.(e-mail: {youky, shijis}@tsinghua.edu.cn)*
** *Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, China. (e-mail: eeqiu@ust.hk)*

Abstract: This paper proposes a randomized incremental algorithm to distributedly compute the least square (LS) estimate of linear systems over sensor networks. By integrating its measurement information, a sensor is randomly activated at every time to incrementally update a diffusion vector, which is also used to recursively estimate the unknown parameters of the system via a temporal average algorithm. Then, the updated diffusion vector is passed to the next activated sensor. The activating process is modeled as an identically and independently distributed process. It is shown that the estimate in each sensor asymptotically converges both in mean and almost surely to the standard LS estimate of the system parameters, which is based on all the sensor information. Simulation is finally included to validate the theoretical results.

Keywords: Distributed estimation, least square, incremental algorithm, sensor network.

1. INTRODUCTION

Recently, research on distributed algorithms over sensor networks has received considerable attention. One of the striking advantages of distributed algorithm lies in locally processing sensor information. Since centralized algorithm utilizes all the sensor information in the fusion center, it is natural to conclude that distributed algorithm might not be as good as its centralized counterpart. This holds only if there is no real-time limitation on the communication network, computational capability and etc, which is unreasonable in the resource limited networks. Actually, it is preferable to adopt distributed algorithm for information processing in the sensor network. This work is concerned with the design of a randomized incremental algorithm to distributedly estimate the unknown parameters of linear systems under the least square estimation error criterion.

Incremental algorithms for distributed optimization have been widely used to minimize a sum of convex functions, and each component function is known only to a particular node of a distributed network [Bertsekas 2010, Nedic and Bertsekas 2001, Johansson et al. 2009]. The key feature of this optimization framework is that each sensor cooperatively estimates a minimizer by using only local information, and is particularly helpful in solving optimization over a large-scale network. Roughly speaking, there are two types of incremental methods for optimization and learning. One is the cyclic incremental subgradient algorithm, where the sensors form a ring structure and sequentially pass the iterate along the ring in clockwise direction. The second one is a noncyclic version using the Markov randomized incremental subgradient method. To achieve the convergence of the incremental iterate to the minimizer, both algorithms have to suitably adjust their step sizes. While under a constant step size, the iterate of the cyclic case converges to a "limit cycle" due to the existence of oscillations of the iterates [Bertsekas 2010].

To overcome this limitation, we note that the oscillations can be averaged out, and propose a temporal *average* algorithm to "maximally" aggregate the information of the historical iterates. In particular, we elaborate this idea in the context of lease square (LS) estimate, where the component function is in a quadratic form, and the aim is to cooperatively estimate the unknown parameters under the LS estimation error criterion.

At every time, a sensor is randomly activated and incrementally update an iterate, which is also named as diffusion vector in this paper. To obtain the LS estimate, each sensor recursively implement an average algorithm to compute the average of all the historical diffusion vectors that have visited this sensor. The activating process of the sensors is modeled as an identically and independently distributed (i.i.d.) process, and the updated diffusion vector is passed to the next activated sensor. Since the diffusion vectors form a randomly switching system, we establish the convergence results from the system point of view, which is substantially different from the approach in Bertsekas [2010], Nedic and Bertsekas [2001], Johansson et al. [2009], and prove the ergodicity of the diffusion vectors, which clearly validates the soundness of the temporal average algorithm by recalling the Birkhoffs Ergodic Theorem [Ash and Doléans-Dade 2000].

The another contribution of this work is on the convergence analysis of the above algorithm. In fact, we show that the estimate of each sensor asymptotically converges both in mean and almost surely to the standard LS estimate, which is computed by using all the sensor information in a centralized approach. It should be noted that there exist other distributed algorithms to compute the LS estimate using only local information. For

⋆ This work was supported by the National Natural Science Foundation of China under grant NSFC 61304038, and the Project-sponsored by SRF for ROCS, SEM.

instance, two parallel consensus-based algorithms [Olfati-Saber and Murray 2004] have been designed to distributedly obtain the LS estimate in Xiao et al. [2005]. However, this method requires to transmit a higher dimension data, and compute the inverse of a square matrix, whose order is of the same as the number of unknown parameters to be estimated. The similar idea has also been pursued in Sayed et al. [2013]. From this perspective, our algorithm is easier to implement and may require less communication load per transmission.

The rest of the paper is organized as follows. The problem under consideration is formally described in Section 2. In Section 3, we explicitly describe our novel distributed algorithms. The convergence analysis is conducted in Section 4. Simulation results are included in Section 5. We draw some concluding remarks in Section 6.

2. PROBLEM FORMULATION

Consider an estimation framework by using N distributed sensors over a network to cooperatively estimate an unknown parameter vector θ. Each sensor takes a noisy measurement as follows

$$y_i = H_i\theta + v_i, i \in \mathcal{V} := \{1, \dots, N\}, \tag{1}$$

where $H_i \in \mathbb{R}^{m\times n}$ is the observation matrix, and $v_i \in \mathbb{R}^m$ is either deterministic or stochastic measurement noise. In a centralized approach, all sensor measurements and observation matrices are transmitted to a remote fusion center via a communication network. The fusion center finally produces an optimal estimate of θ in an appropriate sense.

In this work, we are interested in a distributed approach to compute the least square (LS) estimate of θ, which is obtained via solving the following optimization

$$\widehat{\theta}^* \in \arg\min_{\theta\in\mathbb{R}^n} \frac{1}{2}\sum_{i=1}^N \|y_i - H_i\theta\|^2.$$

By using some standard results on LS techniques [Kailath et al. 2000], it is obvious that if $\sum_{k=1}^N H_k^T H_k$ is positive definite, the LS estimator is exactly expressed as

$$\begin{aligned}\widehat{\theta}^* &= \big(\sum_{i=1}^N H_i^T H_i\big)^{-1}\big(\sum_{i=1}^N H_i^T y_i\big)\\ &= \big(\frac{1}{N}\sum_{i=1}^N H_i^T H_i\big)^{-1}\big(\frac{1}{N}\sum_{i=1}^N H_i^T y_i\big).\end{aligned} \tag{2}$$

To obtain the LS estimator, each sensor is required to send its measurement y_i and observation matrix H_i to the fusion center via a shared network. Clearly, the total dimension of transmitted message of each sensor is $n + n \times m = (1 + m)n$. This may require a high communication capacity of the sensor and increases the energy consumption. It also needs to compute the inverse of a square matrix of order n. This centralized scheme highly relies on the fusion center, and is not very reliable. In practice, it is preferable to *locally* compute the LS estimate at each sensor of the network.

For this purpose, consensus-based estimating algorithms have recently been proposed in the literature, see Xiao et al. [2005] for an example. Observe from (2) that the necessary quantities to compute the LS estimate can be given by two averages

$$\mathcal{H} := \frac{1}{N}\sum_{i=1}^N H_i^T H_i, \text{ and } \mathcal{Y} := \frac{1}{N}\sum_{i=1}^N H_i^T y_i.$$

Thus, it is sufficient to design a distributed algorithm for each sensor to locally compute $\mathcal{H}$ and $\mathcal{Y}$ of the LS estimate. In view of the well established consensus algorithm, a natural idea is to simultaneously implement the following two consensus algorithms at each sensor

$$\begin{aligned}\mathcal{H}_i(t+1) &= \sum_{j\in\mathcal{N}_i} a_{ij}\mathcal{H}_j(t), \mathcal{H}_i(0) = H_i^T H_i,\\ \mathcal{Y}_i(t+1) &= \sum_{j\in\mathcal{N}_i} a_{ij}\mathcal{Y}_j(t), \mathcal{Y}_i(0) = H_i^T y_i,\\ \widehat{\theta}_i(t) &= (\mathcal{H}_i(t))^+\mathcal{Y}_i(t),\end{aligned}$$

where the superscript M^+ denotes the Moore-Penrose pseudoinverse [Horn and Johnson 1985] of matrix M, and $a_{ij} > 0$ if and only if sensor j can send information to sensor i, otherwise $a_{ij} = 0$.

If the communication topology formed by the sensors is connected and undirected, it follows from Olfati-Saber and Murray [2004] that $\lim_{t\to\infty}\mathcal{H}_i(t) = \mathcal{H}$ and $\lim_{t\to\infty}\mathcal{Y}_i(t) = \mathcal{Y}$. Then, each sensor can easily obtain the LS estimate, e.g.,

$$\lim_{t\to\infty}\widehat{\theta}_i(t) = (\lim_{t\to\infty}\mathcal{H}_i(t))^+(\lim_{t\to\infty}\mathcal{Y}_i(t)) = \widehat{\theta}^*.$$

Again, this algorithm requires to exchange the messages of $\mathcal{H}_i(t)$ and $\mathcal{Y}_i(t)$ at each transmission, which has a dimension of $(1 + m)\cdot n$, and compute the inverse (or Moore-Penrose pseudoinverse if necessary) of $\mathcal{H}_i(t)$.

The objective of this paper is to design a novel diffusion-based algorithm over the network to distributedly compute the LS estimate, which is easier to implement and requires less communication load at each transmission than that of the above algorithms, and rigorously establish its asymptotic convergence property.

3. RANDOMIZED INCREMENTAL LS

Motivated by the limitation of the consensus-based algorithm, a novel diffusion-based algorithm over networks is now proposed. At time t, a sensor, indexed as $s(t) \in \mathcal{V}$, is activated and it receives a *diffusion* vector $\widehat{\theta}_0(t)$ (cf. (3)) from the previously activated sensor. Sensor $s(t)$ incorporates its measurement information to incrementally update $\widehat{\theta}_0(t)$ by the following fusion algorithm

$$\begin{aligned}\widehat{\theta}_0(0) &= 0,\\ \widehat{\theta}_0(t+1) &= \widehat{\theta}_0(t) + \alpha\cdot H_{s(t)}^T(y_{s(t)} - H_{s(t)}\widehat{\theta}_0(t)),\end{aligned} \tag{3}$$

where $\alpha > 0$ is the adjustable step size, and is to be designed in the sequel.

Next, the updated $\widehat{\theta}_0(t+1)$ is passed from sensor $s(t)$ to next sensor $s(t+1)$. At time $t+1$, sensor $s(t+1)$ updates $\widehat{\theta}_0(t+1)$ in a similar way by using its own measurement information. By repeating this fashion, it is obvious that $\widehat{\theta}_0(t)$ will be circulated across the networks and visits a sensor at each time.

Note that under a constant step size $\alpha > 0$, it is usually impossible to achieve the convergence of $\widehat{\theta}_0(t)$ for any diffusion process $s(t)$. Otherwise, let $\lim_{t\to\infty}\widehat{\theta}_0(t) = \widehat{\theta}_0(\infty)$. Taking limits on both sides of (3), it follows that $H_i^T y_i = H_i^T H_i\widehat{\theta}_0(\infty)$ for all $i \in \mathcal{V}$, which usually does not hold.

As an initial attempt, we make the following assumption on the diffusion process in this paper.

Assumption 1. The activating process $s(t)$ is assumed to be an identically and independently distributed (i.i.d.) process.

Under the above case, the diffusion vector asymptotically converges in mean to the LS estimate.

Lemma 2. Select a positive $\alpha < 1/\lambda_{\max}(\mathcal{H})$, where $\lambda_{\max}(\mathcal{H})$ is a maximum eigenvalue of $\mathcal{H}$ in magnitude. Then, it holds that

$$\lim_{t\to\infty} \mathbb{E}[\widehat{\theta}_0(t)] = (\alpha\mathcal{H})^{-1}(\alpha\mathcal{Y}) = \widehat{\theta}^*.$$

Proof. By taking expectation on both sides of (3), it follows that

$$\begin{aligned}\mathbb{E}[\widehat{\theta}_0(t+1)] &= (I-\alpha\mathcal{H})\mathbb{E}[\widehat{\theta}_0(t)] + \alpha\mathcal{Y} \\ &= \sum_{i=0}^{t+1}(I-\alpha\mathcal{H})^i(\alpha\mathcal{Y}).\end{aligned} \tag{4}$$

Note that $\lambda_{\max}(I-\alpha\mathcal{H}) < 1$, the rest of the proof is straightforward. ■

By Lemma 2, it is sufficient to design an estimate at each sensor to compute the ensemble average $\mathbb{E}[\widehat{\theta}_0(t)]$, which asymptotically converges to the LS estimate. To this end, each sensor constructs an estimate of θ by taking an average of all the diffusion vectors that have visited the sensor, i.e., sensor i forms an estimate $\widehat{\theta}_i(t)$ of θ by computing the following average,

$$\widehat{\theta}_i(t) = \frac{1}{|\{j \le t | s(j) = i\}|} \sum_{k\in\{j\le t|s(j)=i\}} \widehat{\theta}_0(k) \tag{5}$$

where $|A|$ returns the cardinality of set A.

Under Assumption 1, $\widehat{\theta}_0(t)$ is an ergodic process. By the Birkhoffs Ergodic Theorem [Ash and Doléans-Dade 2000], the temporal average of $\widehat{\theta}_0(t)$ will converge almost surely to its ensemble average. This forms the key basis of the convergence analysis in the next section. It is worthy emphasizing that the estimate in (5) only takes a temporal average of a subset of $\{\widehat{\theta}_0(t), t \ge 0\}$, and the Birkhoffs Ergodic Theorem can not be directly used.

Moreover, the average algorithm (5) lacks a recursive form. To facilitate the implementation, we rewrite the average algorithm in a recursive form. Particularly, suppose that $s(t) = i$, sensor i updates its estimate of the unknown as follows

$$\begin{aligned} m_i(t+1) &= m_i(t) + 1, \\ \widehat{\theta}_i(t+1) &= \frac{m_i(t)}{m_i(t+1)}\widehat{\theta}_i(t) + \frac{1}{m_i(t+1)}\widehat{\theta}_0(t), \end{aligned} \tag{6}$$

and for other sensors, they evolve in an open loop by letting

$$\begin{aligned} m_j(t+1) &= m_j(t), \\ \widehat{\theta}_j(t+1) &= \widehat{\theta}_j(t), \forall j \ne i, \end{aligned} \tag{7}$$

where all of the above quantities are initialized as zero or zero vector, i.e., $m_i(0) = 0$ and $\widehat{\theta}_i(0) = 0$ for all $i \in \mathcal{V}$.

Remark 3. (a) In comparison, the proposed algorithm only requires the activated sensor to diffuse an n-dimension message of $\widehat{\theta}_0(t)$ across the network. It does not require to compute the inverse of any matrix.

(b) The purpose of taking a temporal average in (5) is to maximally aggregate all sensors information on the unknown via the diffusion vectors.

(c) In (3), we consider the randomization on the sensor node. In Ravazzi et al. [2013], the method of using randomization on the network link is proposed.

4. CONVERGENCE ANALYSIS

In this section, we prove that the estimate $\widehat{\theta}_i(t)$ of each sensor will converge both in mean and almost surely to the LS estimate $\widehat{\theta}^*$ as the time t goes to infinity under Assumption 1, which shows the effectiveness of the proposed algorithm for each sensor. The convergence result is formally stated below.

Theorem 4. Suppose that $\sum_{i=1}^N H_i^T H_i$ is positive definite. Under Assumption 1, there exists a sufficiently small positive α^* such that for any positive $\alpha < \alpha^*$ and $i \in \mathcal{V}$ in (6), it holds that

(a) $\lim_{t\to\infty} \widehat{\theta}_i(t) = \widehat{\theta}^*$ almost surely.

(b) $\lim_{t\to\infty} \mathbb{E}[\widehat{\theta}_i(t)] = \widehat{\theta}^*$ where $\mathbb{E}[\cdot]$ is taken with respect to the process $\{s(t)\}$.

Remark 5. In the cyclic incremental algorithm [Nedic and Bertsekas 2001], i.e. $s(t) = t - \lfloor t/m \rfloor + 1$ in (3) where $\lfloor\cdot\rfloor$ is the standard floor function, it was proved that for sufficiently small $\alpha > 0$, there is a limiting point θ_i depending on α such that $\lim_{t\to\infty} \widehat{\theta}_0(tm+i) = \theta_i$ and $\lim_{\alpha\to 0}\theta_i = \theta$ for all $i \in \mathcal{V}$. Since $\theta_i \ne \theta_j$ whenever $i \ne j$, the estimate of every sensor usually does not converge to a same value but a limit cycle.

One may attempt to cancel out the oscillations of $\widehat{\theta}_0(t)$ by designing a temporal average algorithm as this paper. However, this usually can not guarantee the convergence to the exact LS estimate $\widehat{\theta}^*$ under a constant α. To the best of our knowledge, the convergence to the LS estimate under a constant α has not been established [Bertsekas 2010, Nedic and Bertsekas 2001, Johansson et al. 2009].

Example 6. We use a simple example to illustrate the advantages of the randomization over the cyclic incremental algorithm.Let $N = 2$ and $H_1 = H_2 = 1$ in (1). Then, the cyclic incremental algorithm is given by

$$\theta_0(t+1) = \theta_0(t) + \alpha(y(t) - \theta_0(t)), \tag{8}$$

where $y(t) = y_1$ if t is even, and otherwise $y(t) = y_2$. If $0 < \alpha < 1$, one can readily show that

$$\begin{aligned}\lim_{t\to\infty}\theta_0(2t) &= \frac{(\alpha-\alpha^2)y_1 + \alpha y_2}{1-(1-\alpha)^2} \ne \frac{y_1+y_2}{2}, \\ \lim_{t\to\infty}\theta_0(2t+1) &= \frac{(\alpha-\alpha^2)y_2 + \alpha y_1}{1-(1-\alpha)^2} \ne \frac{y_1+y_2}{2},\end{aligned}$$

which means that neither $\theta_0(2t)$ nor $\theta_0(2t+1)$ converge to the LS estimate. If $y_1 \ne y_2$, then it is clear that $\lim_{t\to\infty}\theta_0(2t) \ne \lim_{t\to\infty}\theta_0(2t+1)$. Thus, $\theta_0(t)$ converges to a limit cycle.

However, if $H_1 \ne H_2$ and $0 < \alpha < \max\{1/H_1^2, 1/H_2^2\}$, it follows that

$$\lim_{t\to\infty}\frac{1}{t}\sum_{i=1}^{t}\widehat{\theta}_0(i) = \frac{(2-\alpha H_2^2)H_1 y_1 + (2-\alpha H_1^2)H_2 y_2}{2(H_1^2 + H_2^2 - \alpha H_1^2 H_2^2)}. \tag{9}$$

This implies that the temporal average of iterates of the cyclic incremental algorithm does not converges to the LS estimate except that $H_1^2 = H_2^2$.

The proof of Theorem 4 depends on the following lemmas.

Lemma 7. (Diaconis and Freedman [1999]) Consider the Markov chain $x(t)$ generated by the following iterations

$$x(t+1) = A(t+1)x(t) + B(t+1), \tag{10}$$

with $(A(t), B(t))$ being an i.i.d. sequence with appropriate dimensions. Suppose that

$$\mathbb{E}[\log^+ \|A(t)\|] < \infty \text{ and } \mathbb{E}[\log^+ \|B(t)\|] < \infty,$$

where $\log^+(x) = \max\{\log(x), 0\}$ for any $x > 0$. The infinite random sum

$$B(1) + \sum_{t=1}^{\infty} (A(1)\cdots A(t))B(t+1) \tag{11}$$

converges almost surely to a finite limit if and only if

$$\inf_{t>0} \frac{1}{t}\mathbb{E}[\log \|A(1)\cdots A(t)\|] < 0.$$

Moreover, the distribution of (11) is the unique invariant distribution of the Markov chain $x(t)$.

Lemma 8. (Furstenberg et al. [1960]) Suppose that

$$\mathbb{E}[\log^+ \|A(1)\|] < \infty,$$

it almost surely holds

$$\lim_{t\to\infty} \frac{1}{t}\mathbb{E}[\log \|A(1)\ldots A(t)\|] = \lim_{t\to\infty} \frac{1}{t} \log \|A(1)\ldots A(t)\|$$

where $A(t)$ is a stationary and ergodic sequence.

Lemma 9. Suppose that $A(t)$ is a stationary and ergodic sequence, and $\mathbb{E}[\|A(1)\|] < \infty$. For any $\epsilon > 0$, the following limit exists

$$\lambda(\epsilon) := \lim_{t\to\infty} \frac{1}{t} \log \|(I - \epsilon A(t))\ldots(I - \epsilon A(1))\|. \tag{12}$$

Let $\mu(\cdot)$ be a matrix measure defined as

$$\mu(A) = \lim_{\epsilon\to 0^+} \frac{\|I + \epsilon A\| - 1}{\epsilon}.$$

Then, it holds that

$$\limsup_{\epsilon\to 0^+} \frac{\lambda(\epsilon)}{\epsilon} \le \mu(-\mathbb{E}[A(1)]).$$

Proof of Theorem 4:

(a) Select a positive $\alpha < 1/\lambda_{\max}(\mathcal{H})$, and let $A(t+1) = I - \alpha \cdot H_{s(t)}^T H_{s(t)}$ and $B(t+1) = \alpha \cdot H_{s(t)}^T y_{s(t)}$. It follows from (3) that

$$\widehat{\theta}_0(t+1) = A(t+1)\widehat{\theta}_0(t) + B(t+1), \widehat{\theta}_0(0) = 0. \tag{13}$$

Since $s(t)$ is an i.i.d. process and uniformly distributed, the sequence of $H_{s(t)}^T H_{s(t)}$ is also an i.i.d. process, and

$$\mathbb{E}[H_{s(t)}^T H_{s(t)}] = \frac{1}{N}\sum_{i=1}^{N} H_i^T H_i < \infty.$$

By Lemma 9, the following Lyapunov exponent is well defined

$$\lambda(\alpha) := \lim_{t\to\infty} \frac{1}{t} \log \|A(1)\ldots A(t)\|, \tag{14}$$

and

$$\begin{aligned}\limsup_{\alpha\to 0^+} \frac{\lambda(\alpha)}{\alpha} &\le \mu(-\mathbb{E}[H_{s(t)}^T H_{s(t)}]) = \mu(-\mathcal{H}) \\ &= -\lambda_{\min}(\mathcal{H}),\end{aligned} \tag{15}$$

where $\lambda_{\min}(\mathcal{H})$ is a minimum eigenvalue of $\mathcal{H}$ in magnitude.

Since $\sum_{i=1}^N H_i^T H_i$ is positive definite, it is obvious that $\lambda_{\min}(\mathcal{H}) > 0$. Jointly with (15), there exists a positive α_0 such that $\lambda(\alpha) \le -\alpha\lambda_{\min}(\mathcal{H}) < 0$ for all $\alpha < \alpha_0$. In what follows, α is selected to be a positive number that is strictly less than $\alpha_1 = \min\{\alpha_0, 1/\lambda_{\max}(\mathcal{H})\}$. Note that $\mathbb{E}[\log^+ \|A(1)\|] < \infty$, it follows from Lemma 8 that

$$\begin{aligned}\inf_{t>0} \frac{1}{t}\mathbb{E}[\log \|A(1)\ldots A(t)\|] &\le \lim_{t\to\infty} \frac{1}{t}\mathbb{E}[\log \|A(1)\ldots A(t)\|] \\ &= \lim_{t\to\infty} \frac{1}{t} \log \|A(1)\ldots A(t)\| = \lambda(\alpha) < 0 \text{ almost surely.}\end{aligned}$$

Since $\mathbb{E}[\log^+ \|B(1)\|] < \infty$, it follows from Lemma 7 that $\widehat{\theta}_0(t)$ converges almost surely to (11). Together with (11), it implies that

$$\begin{aligned}&\mathbb{E}[B(1) + \sum_{t=1}^{\infty} (A(1)\cdots A(t))B(t+1)] \\ &= \sum_{t=0}^{\infty} (I - \alpha\mathcal{H})^t(\alpha\mathcal{Y}) = \widehat{\theta}^*.\end{aligned} \tag{16}$$

Define an indicator function

$$\xi_i(t) = \begin{cases} 1, \text{ if } s(t) = i. \\ 0, \text{ otherwise.}\end{cases} \tag{17}$$

Then, $m_i(t) = \sum_{k=1}^t \xi_i(t)$. By the Strong Law of Large Numbers [Ash and Doléans-Dade 2000], it follows that

$$\lim_{t\to\infty} \frac{m_i(t)}{t} = \mathbb{P}\{\xi_i(t) = 1\} = \frac{1}{|\mathcal{V}|} \tag{18}$$

with probability one. For $m_i(t) > 0$, it follows from (6) that

$$\widehat{\theta}_i(t) = \frac{1}{m_i(t)} \sum_{j=1}^{t} \xi_i(j)\widehat{\theta}_0(j). \tag{19}$$

Consider the following auxiliary process

$$\begin{bmatrix}\xi_i(t+1) \\ z(t+1)\end{bmatrix} = \begin{bmatrix}0 & 0 \\ 0 & A(t+1)\end{bmatrix}\begin{bmatrix}\xi_i(t) \\ z(t)\end{bmatrix} + \begin{bmatrix}1_{\{s(t+1)=i\}} \\ B(t+1).\end{bmatrix} \tag{20}$$

Here $z(0)$, which is independent of $\widehat{\theta}_0(0)$, is initialized as a random variable with the same distribution as that of (11), and $1_{\{s(t+1)=i\}}$ is an indicator function, which is one if $s(t+1) = i$ and zero, otherwise. Since $s(t)$ is an i.i.d. process and $\theta_0(t)$ converges almost surely to (11), it is not difficult to verify that $[\xi_i(t), z(t)^T]^T$ is a stationary and ergodic process [Ash and Doléans-Dade 2000].

By the Birkhoff's Ergodic theorem [Ash and Doléans-Dade 2000], it follows that

$$\begin{aligned}\lim_{t\to\infty} \frac{1}{t}\sum_{j=1}^{t} \xi_i(j)z(j) &= \mathbb{E}[\xi_i(1)z(1)] \\ &= \mathbb{E}[\xi_i(1)]\mathbb{E}[z(1)] \\ &= \widehat{\theta}^*/|\mathcal{V}|,\end{aligned} \tag{21}$$

where the second equality is due to that $s(t)$ is independent of $z(t)$ for all $t \ge 0$, and the last equality follows from (16).

Next, we shall prove that

$$\lim_{t\to\infty} \frac{1}{t}\sum_{j=1}^{t} \xi_i(j)\|z(j) - \widehat{\theta}_0(j)\| = 0 \text{ almost surely.} \tag{22}$$

Given a positive $\epsilon < |\lambda(\alpha)|$ and let $\beta := \epsilon + \lambda(\alpha) < 0$, it follows from (14) that there exists a sufficiently large t_0 such that for any $t > t_0$. Then, we obtain that

$$\|A(t)\ldots A(1)\| \le \exp(\beta t). \tag{23}$$

Since $z(0)$ has the same distribution as that of (11), it follows that

$$\mathbb{E}[\|z(0)\|] = \mathbb{E}[\|B(1) + \sum_{t=1}^{\infty}(A(1)\cdots A(t))B(t+1)\|].$$

Let $M_0 := \max_{i\in\mathcal{V}} \|H_i^T y_i\|$, then for all $t > t_0$, we obtain

$$\begin{aligned}\mathbb{E}[\|z(0)\|] &\le M_0(\sum_{j=1}^{\infty}\mathbb{E}[\|A(1)\ldots A(j)\|]+1)\\ &\le M_0(\sum_{j=1}^{t_0}\mathbb{E}[\|A(1)\ldots A(j)\|] + \frac{\exp(\beta)}{1-\exp(\beta)}\\ &:= M_1 < \infty,\end{aligned}$$

where the second inequality follows from (23).

For any $t > t_0$ and given a positive η, it follows from Chebyshevs inequality [Ash and Doléans-Dade 2000] that

$$\begin{aligned}\mathbb{P}\{\|\widehat{\theta}_0(t) - z(t)\| \ge \eta^t\} &\le \frac{\mathbb{E}[\|\widehat{\theta}_0(t) - z(t)\|]}{\eta^t}\\ \le \frac{\mathbb{E}[\|z(0)\|]}{\eta^t}\mathbb{E}[\|A(t)\ldots A(1)\|] &\le \frac{M_1\exp(\beta t)}{\eta^t}.\end{aligned}$$

Select any $\eta > \exp(\beta)$ and $\eta < 1$, one can easily prove that

$$\sum_{t=1}^{\infty}\mathbb{P}\{\|\widehat{\theta}_0(t) - z(t)\| \ge \eta^t\} < \infty.$$

Together with Borel-Cantelli Lemma [Ash and Doléans-Dade 2000], it holds with probability one that for sufficiently large t, then $\|\theta_0(t) - z(t)\| < \eta^t$. This obviously implies that $\lim_{t\to\infty}\|\theta_0(t) - z(t)\| = 0$ almost surely. It is clear from (18) that $m_i(t)$ tends to infinity with probability one as t goes to infinity. Together with Toeplitz Lemma [Ash and Doléans-Dade 2000], it follows that

$$\lim_{t\to\infty}\frac{1}{m_i(t)}\sum_{j=1}^{t}\xi_i(j)\|z(j) - \widehat{\theta}_0(j)\| = 0 \text{ almost surely.}$$

Then, (22) follows easily, which together with (21) implies that

$$\lim_{t\to\infty}\frac{1}{t}\sum_{j=1}^{t}\xi_i(j)\widehat{\theta}_0(t) = \widehat{\theta}^*/|\mathcal{V}| \text{ almost surely.}$$

Jointly with (18) and (19), it follows that $\lim_{t\to\infty}\widehat{\theta}_i(t) = \widehat{\theta}^*$ almost surely.

(b) It is sufficient to prove the *uniform integrability* of $\widehat{\theta}_i(t)$ since together with part (a), it follows from Theorem 6.5.2 in Ash and Doléans-Dade [2000] that

$$\lim_{t\to\infty}\mathbb{E}[\widehat{\theta}_i(t)] = \mathbb{E}[\lim_{t\to\infty}\widehat{\theta}_i(t)] = \widehat{\theta}^*. \tag{24}$$

In view of (19), we obtain that for $m_i(t) > 0$ [1], then

$$\widehat{\theta}_i(t) = \sum_{j=1}^{t}\left(\frac{\xi_i(j)}{m_i(t)}\right)\widehat{\theta}_0(j). \tag{25}$$

For $1 < p < \infty$, the p-th moment of a random vector x is defined by $\|x\|_p = (\mathbb{E}[\|x\|^p)^{1/p}$. It is known that $\|\cdot\|_p$ is a norm. Select any $p \in (1,2)$, it follows from the triangle inequality that

$$\|\widehat{\theta}_i(t)\|_p \le \sum_{j=1}^{t}\|\left(\frac{\xi_i(j)}{m_i(t)}\right)\widehat{\theta}_0(j)\|_p. \tag{26}$$

Let $q = 2/p > 1$ and $q' = 2/(2-p)$. By the Holder inequality [Ash and Doléans-Dade 2000], we obtain that

$$\begin{aligned}\|\left(\frac{\xi_i(j)}{m_i(t)}\right)\widehat{\theta}_0(j)\|_p &\le \|\frac{\xi_i(j)}{m_i(t)}\|_{pq'}\cdot\|\widehat{\theta}_0(j)\|_{pq}\\ &\le \sup_{t\ge1}\|\widehat{\theta}_0(t)\|_2\cdot\|\frac{\xi_i(j)}{m_i(t)}\|_{pq'}.\end{aligned}$$

Since $\xi_i(t)$ is i.i.d., then $m_i(t) = \sum_{j=1}^{t}\xi_i(j)$ has a binomial distribution, e.g., $m_i(t) \sim B(t, 1/|\mathcal{V}|)$. One can verify that

$$\sup_{t\ge1}\|\frac{t}{m_i(t)}\|_{pq'} < \infty. \tag{27}$$

Next, we show that $\sup_{t\ge1}\|\widehat{\theta}_0(t)\|_2 < \infty$. To this purpose, define a Lyapunov functional candidate

$$V(t) = \mathbb{E}[\|\widehat{\theta}_0(t)\|^2].$$

By (13), we obtain that

$$\begin{aligned}\mathbb{E}[V(t+1)|\widehat{\theta}_0(t)] &= \widehat{\theta}_0(t)^T\mathbb{E}[A(t+1)]\widehat{\theta}_0(t)\\ &+ 2\widehat{\theta}_0(t)^T\mathbb{E}[A(t+1)B(t+1)] + \mathbb{E}[\|B(t+1)\|^2].\end{aligned} \tag{28}$$

In addition, it is easy to compute that

$$\mathbb{E}[A(t+1)^2] = I - 2\alpha\mathcal{H} + \alpha^2\mathbb{E}[(H_{s(t)}^T H_{s(t)})^2]. \tag{29}$$

For sufficiently small , the right hand side of the above equality will be dominated by the second term. Since $\mathcal{H}$ is positive definite, there exists a positive $\alpha^* < \alpha_1$ such that for any $\alpha < \alpha^*$, it holds that

$$\rho := \lambda_{\max}(\mathbb{E}[A(t+1)^2]) < 1. \tag{30}$$

By Lemma 2, it is clear that

$$\sup_{t\ge1}\|\mathbb{E}[\widehat{\theta}_0(t)]\| < \infty.$$

Since $s(t)$ is i.i.d., there exists a positive constant c such that for all $t \ge 1$,

$$2\cdot\mathbb{E}[\widehat{\theta}_0(t)^T]\mathbb{E}[A(t+1)B(t+1)] + \mathbb{E}[\|B(t+1)\|^2] < c.$$

In light of (28), it yields that

$$V(t+1) \le \rho V(t) + c, \tag{31}$$

which implies that $\mathbb{E}[\|\widehat{\theta}_0(t)\|^2] \le c/(1-\rho) < \infty$.

Together with (26) and (27), it follows that

$$\sup_{t\ge1}\mathbb{E}[\|\widehat{\theta}_i(t)\|^p] < \infty. \tag{32}$$

Since $p > 1$, it follows from Lemma 6.5.6 [Ash and Doléans-Dade 2000] that $\widehat{\theta}_i(t)$ is uniformly integrable. ■

5. SIMULATION

Consider a linear system as follows

$$y_i = H_i\theta + v_i, i \in \{1,\ldots,20\}, \tag{33}$$

where the true parameter vector $\theta = [1,2,1.5]^T$ and v_i is a white Gaussian noise with zero mean and unit variance. The observation matrix H_i is randomly generated from a Gaussian vector, i.e. $H_i \sim \mathcal{N}(0, I_3)$. The step size is set as $\alpha = 0.02$.

[1] Note that $\widehat{\theta}_i(t) = 0$ if $m_i(t) = 0$.

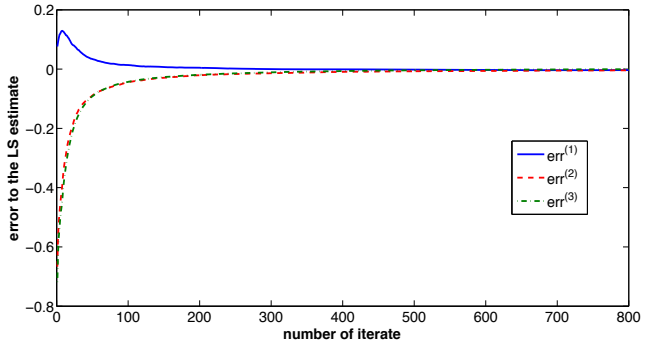

Fig. 1. The transient error of the estimate in a sensor node.

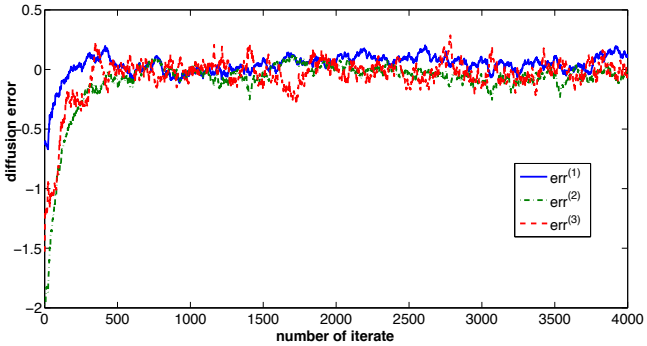

Fig. 2. The ergodicity of the diffusion vector.

To check with the convergence of the estimate at each sensor to the LS estimate, define its transient error by

$$err_i = \widehat{\theta}_i(t) - \widehat{\theta}^*.$$

By Theorem 4, it is expected that $err_i(t)$ will asymptotically converge both in mean and almost surely to a zero vector. We randomly choose a sensor node with an equal probability to illustrate it, and in this simulation, the second sensor node is selected. The behavior of $err_2(t)$ is shown in Fig. 1. It is clear that both theoretical and simulation results are consistent and verify the convergence property of the estimate algorithm. It should be noted that the transient error of all the other sensors will exhibit the similar behavior.

In addition, we also examine the ergodicity of the diffusion vector $\widehat{\theta}_0(t)$, and define the diffusion error

$$err_0(t) = \widehat{\theta}_0(t) - \widehat{\theta}^*.$$

As remarked, the use of the temporal average algorithm is motivated by the ergodicity of the diffusion vector. Thus, it is expected that $err_0(t)$ is an ergodic process with its mean asymptotically converging to zero. This is supported by the simulation result shown in Fig. 2, which validates the soundness of the estimating algorithm.

6. CONCLUSION

Motivated by the ergodicity of randomized incremental algorithms, we have proposed a temporal average algorithm to distributedly compute the lease square (LS) estimate of linear systems. It was rigorously proved that the dis- tributed estimate asymptotically converges both in mean and almost surely to the LS estimate. In the future work, we extend the result to the case that the diffusion vector can only be passed to the neighboring sensors due to the limited communication range.

ACKNOWLEDGEMENT

The authors are grateful to Prof. Roberto Tempo for insightful conversations on the topics of this paper.

REFERENCES

Ash, R. and Doléans-Dade, C. (2000). *Probability and Measure Theory*. Academic Press.

Bertsekas, D.P. (2010). Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. Technical report, MIT, Cambridge, MA.

Diaconis, P. and Freedman, D. (1999). Iterated random functions. *SIAM Review*, 41(1), 45–76.

Furstenberg, H., Kesten, H., et al. (1960). Products of random matrices. *Ann. Math. Statist*, 31(2), 457–469.

Horn, R. and Johnson, C. (1985). *Matrix Analysis*. Cambridge University Press.

Johansson, B., Rabi, M., and Johansson, M. (2009). A randomized incremental subgradient method for distributed optimization in networked systems. *SIAM Journal on Optimization*, 20(3), 1157–1170.

Kailath, T., Sayed, A., and Hassibi, B. (2000). *Linear Estimation*. Prentice Hall Upper Saddle River.

Nedic, A. and Bertsekas, D.P. (2001). Incremental subgradient methods for nondifferentiable optimization. *SIAM Journal on Optimization*, 12(1), 109–138.

Olfati-Saber, R. and Murray, R. (2004). Consensus problems in networks of agents with switching topology and time-delays. *IEEE Transactions on Automatic Control*, 49(9), 1520–1533.

Ravazzi, C., Frasca, P., Tempo, R., and Ishii, H. (2013). Ergodic randomized algorithms and dynamics over networks. *arXiv preprint arXiv:1309.1349*.

Sayed, A.H., Tu, S.Y., Chen, J., Zhao, X., and 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.

Xiao, L., Boyd, S., and Lall, S. (2005). A scheme for robust distributed sensor fusion based on average consensus. In *Fourth International Symposium on Information Processing in Sensor Networks*, 63–70.