Chaos

Adaptive clustering based on element-wised distance for distributed estimation over multi-task networks

Cite as: Chaos **30**, 053116 (2020); https://doi.org/10.1063/1.5141493
Submitted: 05 December 2019 . Accepted: 02 April 2020 . Published Online: 07 May 2020

A. Yuanyuan Zhang, B. Minyu Feng, C. Feng Chen, and D. Jürgen Kurths

Chaos ARTICLE scitation.org/journal/cha

Adaptive clustering based on element-wised distance for distributed estimation over multi-task networks

Cite as: Chaos **30**, 053116 (2020); doi: 10.1063/1.5141493
Submitted: 5 December 2019 · Accepted: 2 April 2020 ·
Published Online: 7 May 2020

View Online Export Citation CrossMark

A. Yuanyuan Zhang,[1,a)] B. Minyu Feng,[1] C. Feng Chen,[1,b)] and D. Jürgen Kurths[2,c)]

AFFILIATIONS

[1] College of Artificial Intelligence, Southwest University, Chongqing 400715, China
[2] Potsdam Institute for Climate Impact Research, 14473 Potsdam, Germany

[a)] **Also at:** College of Electronic and Information Engineering, Southwest University, Chongqing 400715, China and Key Laboratory of Nonlinear Circuits and Intelligent Information Processing.
[b)] **Also at:** College of Electronic and Information Engineering, Southwest University, Chongqing 400715, China, and at the Brain-Inspired Computing and Intelligent Control Key Laboratory, Southwest University, Chongqing 400715, China.
Author to whom correspondence should be addressed: fengchen.uestc@gmail.com
[c)] **Also at:** Department of Physics, Humboldt University, Berlin 12489, Germany.

ABSTRACT

In multitask networks, neighboring agents that belong to different clusters pursue different goals, and therefore arbitrary cooperation will lead to a degradation in estimation performance. In this paper, an adaptive clustering method is proposed for distributed estimation that enables agents to distinguish between subneighbors that belong to the same cluster and those that belong to a different cluster. This creates an appropriate degree of cooperation to improve parameter estimation accuracy, especially for the case where the prior information of a cluster is unknown. In contrast to the static and quantitative threshold that is imposed in traditional clustering methods, we devise a method for real-time clustering hypothesis detection, which is constructed through the use of a reliable adaptive clustering threshold as reference and the averaged element-wise distance between tasks as real-time clustering detection statistic. Meanwhile, we relax the clustering conditions to maintain maximum cooperation without sacrificing accuracy. Simulations are presented to compare the proposed algorithm and some traditional clustering strategies in both stationary and nonstationary environments. The effects of task difference on performance are also obtained to demonstrate the superiority of our proposed clustering strategy in terms of accuracy, robustness, and suitability.

Distributed adaptive algorithms for learning and optimization by networked nodes are effective and have recently become popular in wireless sensor network signal processing and parameter estimation. The objective of the present work is to create reasonable cooperation to improve the accuracy of parameter estimation in a distributed manner, mainly for multitask networks where the prior information of clusters is unknown. This is a challenging task for most methods. Thus, we propose an adaptive clustering method, which is performed by a real-time clustering hypothesis test through an adaptive clustering threshold based on element-wise distance. Following online clustering for efficient cooperation, a distributed estimation algorithm over multitask networks is proposed. Simulations are provided to show the superior performance of our proposed algorithms, which are more suitable, robust, and accurate than previous strategies, regardless of the difference in local optima in stationary and nonstationary environments.

I. INTRODUCTION

In multi-agent networks, distributed adaptive algorithms for learning, inference, adaptation, modeling, and optimization by networked nodes are effective and popular in online supervised learning, reinforcement learning, and signal processing. In a centralized solution for parameter estimation, each agent transmits its local data to a central fusion center for processing. Centralizing all streaming measurements to a single node as the fusion center is a high-risk

task, because single-node failure is nonrobust and lacks scalability; also, it may require large amounts of energy and communication resources.

Performing adaptive estimation in a distributed manner is a more robust and resource-saving method to solve inference problems in an autonomous and collaborative way, which depends on each agent over the whole network. Several strategies have been proposed for distributed information processing over networks, including consensus strategies,[1,2] incremental strategies,[3,4] and diffusion strategies.[5,6] Diffusion strategies have attracted more attention owing to their low power consumption, enhanced adaptation performance, and wider stability ranges when constant step sizes are used to enable continuous learning scalability as well as robustness.[7]

In a single-task network in which a diffusion approach is adopted, each node in the network exchanges information only with neighboring nodes, and processing is distributed among all nodes in the network through global diffusion. In the case where all agents pursue a common unknown target parameter, cooperation through distributed diffusion across the whole network is beneficial to improving the performance of parameter estimation. It is worth noting that adapt-then-combine (ATC) diffusion least mean square (d-LMS) is a diffusion-based adaptive solution of the LMS type to distributed optimization problems.

Recently, distributed adaptive estimation based on the ATC d-LMS algorithm has been studied extensively and in depth in multitask networks for issues such as multitarget location tracking, classification, and clustering.[8–12] Hence, our work here concentrates on the distributed diffusion implementation for multitask network parameter estimation.

In a multitask network, the local optima vary among different clusters. Agents belonging to the same cluster share a common optimum by minimizing the cluster cost function, which aggregates each agent's local cost functions from the same cluster in a distributed manner. Nodes exchange information locally and cooperate only with their subneighbors belonging to the same cluster, without the need for sharing or requiring any useless and task-irrelevant global information.

In previous work, it has usually been assumed that the multitask network is completely clustered, and that each agent obtains prior knowledge of clusters so that differentiated cooperation is possible. However, arbitrary cooperation will lead to estimation performance degradation under the circumstance that nodes belonging to different clusters have various objectives and that prior information on the cluster is unknown. More narrowly, each agent has no prior knowledge about which subneighbors belong to the same cluster as itself. As a result, agents pursuing unique tasks cannot cooperate with all their neighbors directly as with global diffusion transmission in a single-task network. Thus, clustering over multitask networks is indispensable for distributed estimation.

Some common paradigms of traditional strategies have been applied for parameter estimation in this situation, such as the use of an adaptive combination rule to achieve differential cooperation, and, in order to reduce the mean square-deviation (MSD), a probability parameter is introduced to provide a measure of certainty about information fusion in estimation[13–21] with the aim of reducing communication load. Also, some innovative ideas such as bacterial foraging optimization and the chaotic ant swarm approach have been proposed for data clustering.[22,23] To measure the degree of similarity among clusters, the Euclidean distances of the entire parameter vectors between different agents are calculated for comparison with a predefined static threshold.[24] However, these methods not only have the disadvantage of low estimation accuracy, but are also sensitive to the setting of some static clustering threshold. Additionally, there are some limitations in the studies mentioned above, especially with regard to the use of an adaptive combination weight according to the differences among different tasks of all neighbors.[13,14] Even if an algorithm converges toward the Pareto optimum over a multitask network,[25–27] the bias can be large if the distance between local optima is large. A global diffusion strategy has been adopted for multitask network estimation[16] in which modeling is performed on local optima that are not significantly distinct from each other. This strategy provides an improved performance only when the local optima differ only slightly. Conversely, when local optima with significant difference in magnitude are pursued, the accuracy of task estimation is reduced as a result of the inevitable introduction of a large number of useless and misleading task-irrelevant neighbors. In this work, we extend this approach by assuming that the local optima are significantly different in each dimension to remedy the limitations of previous works. We use an adaptive clustering threshold, rather than assuming a static clustering threshold and changing the combination weight across the whole network.

Compared with previous methods, we make the following new contributions in this work. First, to minimize the deviation between estimators and the system's true value[28] and also to take into account the nonstationary environment, we derive a fully distributed adaptive clustering threshold based on element-wise distance for the differences in each dimension. We make a one-step approximation to obtain an approximate optimal value that can provide the basis for determining an accurate clustering threshold and a distributed estimation algorithm capable of adapting and learning from measurement samples. Second, we extend this by taking into consideration the significant differences among local optima to remedy the limitation of previous approaches.[16] In contrast to existing algorithms, we propose an adaptive clustering method that is constructed through a real-time clustering hypothesis detection test using an adaptive clustering threshold as reference for accurate clustering and the averaged element-wise distance between different tasks to further enhance clustering accuracy, instead of adopting a static threshold and using an adaptive combination weight across the network to bring bias into estimation. We also add a relaxation factor to provide a trade-off between clustering accuracy and maximum cooperation for reducing the steady-state misalignment of estimation. Third, we devise a distributed parameter estimation algorithm based on the adaptive clustering method over multitask networks. Finally, we present simulation results to demonstrate the performance of the proposed clustering strategy and the distributed parameter estimation algorithm with the adaptive clustering method over multitask networks. The clustering method developed here is more suitable, robust, and accurate than traditional clustering methods, regardless of the differences in tasks between stationary and nonstationary environments.

The rest of this paper is structured as follows. We construct the system model in Sec. II. In Sec. III, we formulate the multitask

estimation problem, and in Sec. IV, we elaborate the proposed clustering algorithm. We present the simulation results in Sec. V. Finally, this work is summarized in Sec. VI.

II. SYSTEM MODEL

In this section, the multitask basic network model and the data model are described.

A. Networks with multiple clusters

We consider a connected network with a node set $S = \{1, 2, \ldots, N\}$ categorized into s mutually exclusive clusters, which are denoted by $\{\mathcal{C}_n\}_{n=1}^{s}$. Each node k of the network can communicate with its adjacent agents $\mathcal{N}_k$. We denote a real-time cluster including node k by $\mathcal{C}_{k,i} = \mathcal{N}_k \bigcap \mathcal{N}_{k,i}^{+}$, where $\mathcal{N}_{k,i}^{+}$ represents subneighbors with the same objective as node k collected through clustering detection at time i. $\mathcal{N}_{k,i}^{-} \triangleq \mathcal{N}_k \backslash \mathcal{C}_{k,i}$ represents subneighbors with different objectives from node k at time i. However, the subneighbors in the same or different clusters of agents in the initial stage are not clear.

This is illustrated by Fig. 1(a), which gives an example of such a multitask network. A total of 11 nodes belonging to 3 clusters are shown in the figure, with nodes belonging to different clusters being represented by colored circles. Nodes at either end of a solid line are subneighbors of node k in the same cluster, while the dotted lines link subneighbors in different clusters. Specifically, in Fig. 1(a), node k and its subneighbors of the same cluster form $\mathcal{N}_k^{+} = \{l, k, 5, 6\}$, while the subneighbors of a cluster different from that of node k form $\mathcal{N}_k^{-} = \{4\}$. The clusters evolve over time as the proposed dynamics clustering detection is applied to multitask network, and Fig. 1(b) shows the steady-state subnetworks for each cluster after the clustering evolutionary process. It is clear that the whole network is eventually divided into three subnetworks, which represents three different clusters that have evolved through the clustering process.

B. Data model

In this paper, we consider a multitask network environment where different clusters perform different estimation tasks. It is assumed that each agent k is interested in estimating a unique $M \times 1$ unknown optimum weight vector w_k^o. Agents of the same cluster estimate the same optimum

$$w_k^o = w_{\mathcal{C}_n}^o \quad \forall k \in \mathcal{C}_n, \tag{1}$$

where the cluster $\mathcal{C}_n \in \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_s\}$ is in the multitask network, and each node k collects scalar measurements $d_k(i)$ and $1 \times M$ regression data vectors $u_{k,i}$ over successive time instants i. The measurements across all nodes are assumed to be related to a set of unknown $M \times 1$ optimum weight vectors $w^o \triangleq \mathrm{col}\{w_k^o\}_{k=1}^{N}$ via a linear regression model of the form

$$d_k(i) = u_{k,i} w_k^o + v_k(i), \tag{2}$$

where $v_k(i)$ is a zero-mean i.i.d. additive Gaussian measurement noise, which is assumed to be independent of any other signals and to have covariance matrix $R_{v,k} = \sigma_{v,k}^2 I_M$, and w_k^o denotes the local optimum of node k.

III. PROBLEM FORMULATION

There are s clusters $\{\mathcal{C}_n\}_{n=1}^{s}$ in a multitask network. Since agents belonging to different clusters pursue different and irrelevant goals, the objective of nodes in the multitask network is equivalent to the procedure for dealing with a clustered multitask problem through seeking the unique minimizer of the aggregate cluster cost function, which can be written as

$$\underset{\{w_n\}_{n=1}^{s}}{\text{minimize}} \, J^{\text{glob}}(w_1, w_2, \ldots, w_s) \triangleq \sum_{n=1}^{s} J_n^{\mathcal{C}}(w_n). \tag{3}$$

To minimize all cluster cost functions defined by (3), agents need to cooperate only within their clusters, since the objective of each cluster is completely irrelevant, and cooperation with neighbors that belong to different clusters may cause biased effects due to misleading information from these neighbors.[10,13–15] Thus, diffusion strategies can only be applied within each cluster:

$$\underset{w_n}{\text{minimize}} \, J_n^{\mathcal{C}}(w_n) \triangleq \sum_{k \in \mathcal{C}_n} J_k(w_n). \tag{4}$$

However, in the scenario where cluster information is completely unavailable, each node $k \in \mathcal{C}_n$ performs a self-governed estimation task following the principles of distributed optimization, namely, to estimate w_n of the local optima set $w^o \triangleq \mathrm{col}\{w_n\}_{n=1}^{s}$ by seeking the solution for a twice-differentiable local cost function, denoted by $J_k(w_n) \in \mathbb{R}$. Hence, the global cost function can

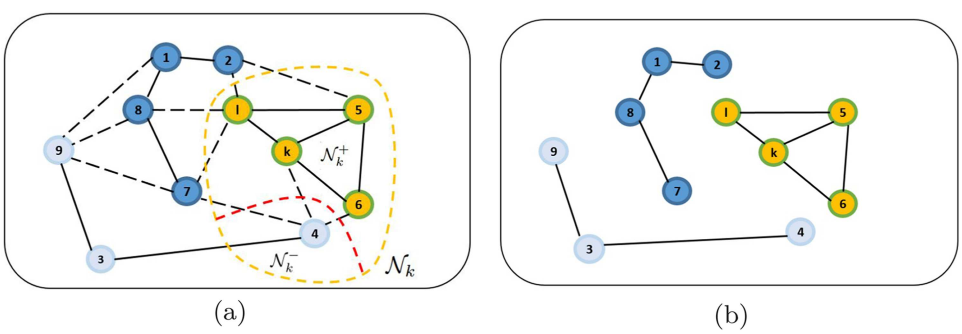

FIG. 1. Multitask network topology: (a) networks with multiple cluster topology; (b) steady-state subnetworks for each cluster after the clustering evolution process.

be decomposed into a single node, which is expressed in the mean-square-error (MSE) form

$$J^{\text{glob}}(w_1, w_2, \ldots, w_s) \triangleq \sum_{n=1}^{s} \sum_{k\in\mathcal{C}_n} J_k(w_n)$$
$$= \sum_{n=1}^{s} \sum_{k\in\mathcal{C}_n} \tfrac{1}{2} E|d_k(i) - u_{k,i} w_{k,i}|^2, \quad (5)$$

for the cluster $\mathcal{C}_n \in \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_s\}$.

Recalling that $\mathcal{C}_{k,i} = \mathcal{N}_k \bigcap \mathcal{N}_{k,i}^{+}$, and in order to minimize the global cost function that aggregates all agents' local cost functions though the ATC d-LMS algorithm, we solve the global optimization problem by applying the gradient descent method to each single node's optimization problem in a distributed manner, which can be expressed as

$$\begin{cases} \varphi_{k,i} = w_{k,i-1} - \mu_k \nabla J_k(w_{k,i-1}), \\ w_{k,i} = \sum\limits_{l\in\mathcal{N}_{k,i}^{+}} c_{lk,i}\varphi_{l,i}, \end{cases} \quad (6)$$

by applying ATC d-LMS in a multitask network. The procedure in (6) consists of an incremental update followed by a diffusion update as a single-task network. The first adaptive step uses an iterative steepest-descent solution to minimize each single agent's cost function, adding new innovation to $\varphi_{k,i}$ at each iteration. The second step represents a convex combination of estimates from LMS filters fed by spatially distinct data from $l \in \mathcal{N}_{k,i}^{+}$, the subneighbors with the same objective. That is to say, it performs cooperation among subneighbors of the same cluster. Following these two steps, $w_{k,i}$ gradually approaches w_k^o. Here, $\mu_k > 0$ denotes the step size and $w_{k,i}$ is an instantaneous estimate of w_k^o at instant i. $\nabla J_k(w_{k,i-1})$ is given by[3]

$$\nabla J_k(w_{k,i-1}) = R_{u,k} w_{k,i-1} - R_{du,k}. \quad (7)$$

Substituting (7) into (6) leads to the steepest-decent step:

$$w_{k,i} = w_{k,i-1} - \mu_k (R_{u,k} w_{k,i-1} - R_{du,k}). \quad (8)$$

Since there is no prior knowledge about the second-order statistical data $\{R_{du,k}, R_{u,k}\}$, we replace this by local instantaneous approximations,[5,6] such as those of LMS type $R_{u,k} \approx u_{k,i}^T u_{k,i}$, $R_{du,k} \approx d_k(i) u_{k,i}^T$. Then, (6) can be further simplified to

$$\begin{cases} \varphi_{k,i} = w_{k,i-1} + \mu_k u_{k,i}^T (d_k(i) - u_{k,i} w_{k,i-1}), \\ w_{k,i} = \sum\limits_{l\in\mathcal{N}_{k,i}^{+}} c_{lk,i}\varphi_{l,i}. \end{cases} \quad (9)$$

Put another way, as an active executor in a multitask network, each node executes the ATC algorithm and obtains a local estimator by first clustering over all neighbors with the intermediate estimator $\varphi_{k,i}$ and then performing cooperation among subneighbors of the same cluster. Thus, the global optimization problem can be decomposed into several local optimization problems in a distributed manner through our proposed adaptive clustering and estimation strategy in a multitask network.

IV. PROPOSED ALGORITHM

If it is assumed that each agent has no prior information about the clusters, then clustering become a crucial issue for each agent to achieve its goal of cooperation within a certain range.

In this section, a real-time clustering detector $T_{lk,i}$ is constructed between $l \in \mathcal{N}_k$ and node k at time i, which will lead to the evolution of clusters over time:

$$\mathcal{N}_{k,i}^{+} \triangleq \{l \in \mathcal{N}_k \mid T_{lk,i} = \theta_{lk,i} \le \lambda_k \gamma_{k,i}\}. \quad (10)$$

In this work, we aim at clustering (i.e., collecting) subneighbors l of node k into $\mathcal{N}_{k,i}^{+}$, where $\mathcal{N}_{k,i}^{+}$ denotes the subneighbors belonging to the same cluster as node k at time i. $\gamma_{k,i}$ represents an adaptive clustering threshold as reference, and λ_k is some constant, defined as a relaxation factor that relaxes the clustering conditions while ensuring cooperation with subneighbors. $\theta_{lk,i}$ denotes the element-wise clustering detection statistic.

To construct the clustering detector $T_{lk,i}$, an approximate optimal reference vector $\widehat{w}_{k,i}^o$ is obtained first in Sec. IV A, and this provides a foundation for constructing the accurate clustering threshold. Then, an adaptive clustering threshold is devised in Sec. IV B and an element-wise clustering detection statistic $\theta_{lk,i}$ in Sec. IV C.

The transient weight error vector of the approximate optimal reference vector $\widehat{w}_{k,i}^o$ can be defined as

$$\Delta\widehat{w}_{k,i} = \widehat{w}_{k,i}^o - w_k^o. \quad (11)$$

where $\widehat{w}_{k,i}^o$ is each agent k's local approximate optimal reference vector for clustering, and $\varphi_{l,i}$ is the observable intermediate estimate of k's neighbors $l \in \mathcal{N}_k$ that is received by k. On the premise that any agent l belongs to the subneighbors of agent k that are in the same cluster as k, i.e., $l \in \mathcal{C}_{k,i}$, then $\varphi_{l,i}$ should be close to agent k's local approximate optimal reference vector $\widehat{w}_{k,i}^o$ and bounded to a first approximation. For the sake of simplicity, let $\lambda_k = 1$, so that $\varphi_{l,i}$ is bounded as

$$\widehat{w}_{k,i}^o - \sqrt{M\cdot\gamma_{k,i}} \cdot \mathbb{1}^T \preceq \varphi_{l,i} \preceq \widehat{w}_{k,i}^o + \sqrt{M\cdot\gamma_{k,i}} \cdot \mathbb{1}^T. \quad (12)$$

We can then write the relationship between $\varphi_{l,i}$ and $\widehat{w}_{k,i}^o$ as follows:

$$\varphi_{l,i} = \widehat{w}_{k,i}^o + \xi_{lk,i}, \qquad l \in \mathcal{N}_{k,i}^{+}. \quad (13)$$

$\xi_{lk,i} = [\xi_{lk,i}^{(1)}, \xi_{lk,i}^{(2)}, \ldots, \xi_{lk,i}^{(M)}]^T$ represents the difference between $\varphi_{l,i}$ and $\widehat{w}_{k,i}^o$. It can be seen from the above analysis that $\xi_{lk,i}$ is an M-dimensional bounded error vector with an average element length, called the element-wise distance $\theta_{lk,i}$, that is bounded by

$$\theta_{lk,i} = \|\widehat{w}_{k,i}^o - \varphi_{l,i}\|^2/M \le \gamma_{k,i}; \quad (14)$$

that is, $\|\xi_{lk,i}\|^2/M \le \gamma_{k,i}$.

A. One-step approximation

To reduce the MSD bias that results from the cooperation of nodes executing various estimation tasks, an approximate optimal vector $\widehat{w}_{k,i}^o$ can be obtained by using the intermediate estimator $\varphi_{k,i}$ from data diffusion at time i. In particular, we make a local one-step approximation based on $\varphi_{k,i}$, thereby obtaining $\widehat{w}_{k,i}^o$ as node k's

approximate optimal vector at time i, which acts as a more reliable reference than $\varphi_{k,i}$:

$$\begin{aligned}\widehat{w}_{k,i}^{o} &= \varphi_{k,i} - \mu_k \nabla_k J_k(w)\big|_{w=\varphi_{k,i}} \\ &\approx \varphi_{k,i} - \mu_k \widehat{\nabla J_k}(w)\big|_{w=\varphi_{k,i}} \\ &= \varphi_{k,i} + \mu_k u_{k,i}^T (d_k(i) - u_{k,i}\varphi_{k,i}). \end{aligned} \tag{15}$$

In (15), $\widehat{w}_{k,i}^{o}$ results from performing a further gradient descent about the local cost function based on the intermediate estimator $\varphi_{k,i}$. We have made the approximation of replacing $\nabla_k J_k(w)$ at $\varphi_{k,i}$ by the instantaneous value $-(u_{k,i}^T(d_k(i) - u_{k,i}\varphi_{k,i}))$.

Since the neighbors of node k are composed of nodes performing a variety of different tasks, node k performs the d-LMS algorithm with information exchange only among $l \in \mathcal{N}_{k,i}^{+}$ after we collect a set of subneighbors of node k in (9). To minimize the transient weight error vector as much as possible, we have narrowed the range of information fusion. However, the information $\varphi_{k,i}$ coming from $l \in \mathcal{N}_{k,i}^{+}$ still gives rise inevitably to an estimation bias in the combination step. Through this step, we reduce intermediate estimator deviation as much as possible in terms of decreasing the local cost.

As time $i \to \infty$, $\widehat{w}_{k,i}^{o}$ approaches the cluster objective w_k^o, which lays a foundation for constructing an accurate clustering threshold.

B. Adaptive element-wise clustering threshold

In this subsection, we calculate the clustering threshold in an element-wise sense, which ensures that the optima of agents in the same cluster are similar in each dimension. Since the local optimum parameter is a multidimensional vector that has real physical meaning in each dimension, to make sense in practical applications, we should measure the similarity of all the dimensions.

First, we substitute (13) into the combination step in (9), and the estimator can then be rewritten as

$$\begin{aligned} w_{k,i} &= \sum_{l\in\mathcal{N}_{k,i}^{+}} c_{lk,i}\varphi_{l,i} \\ &= \sum_{l\in\mathcal{N}_{k,i}^{+}} c_{lk,i}(\widehat{w}_{k,i}^{o} + \xi_{lk,i}) \\ &= \widehat{w}_{k,i}^{o} + \sum_{l\in\mathcal{N}_{k,i}^{+}} c_{lk,i}\xi_{lk,i} \\ &= \widehat{w}_{k,i}^{o} + \mathrm{C}_i\Theta_{k,i}, \end{aligned} \tag{16}$$

where $\Theta_{k,i} = [\xi_{1k,i}, \xi_{2k,i}, \ldots, \xi_{Nk,i}]^T$. Subtracting w_k^o from both sides of (13), we have

$$\Delta w_{k,i} = \Delta\widehat{w}_{k,i} + \mathrm{C}_i\Theta_{k,i}. \tag{17}$$

We then obtain the transient squared deviation between the d-LMS estimator $w_{k,i}$ and $\widehat{w}_{k,i}^{o}$ as

$$\|\Delta w_{k,i}\|^2 = \|\Delta\widehat{w}_{k,i}\|^2 + \|\mathrm{C}_i\Theta_{k,i}\|^2 + 2\Delta\widehat{w}_{k,i}{}^T\mathrm{C}_i\Theta_{k,i}. \tag{18}$$

To minimize the diffusion estimation bias, we first impose a strict condition to ensure that the transient d-LMS estimator $w_{k,i}$ approaches the reliable approximate optimal vector $\widehat{w}_{k,i}^{o}$ as closely as possible. From $\Delta w_{k,i} = w_{k,i} - w_k^o$ and (11), and if $w_{k,i}$ is optimal, then $w_{k,i}$ is closer to the system optimum w_k^o, so the strict condition $\|\Delta w_{k,i}\|^2 \le \|\Delta\widehat{w}_{k,i}\|^2$ must hold. We then get the following relationship:

$$\|\mathrm{C}_i\Theta_{k,i}\|^2 \le 2\Delta\widehat{w}_{k,i}{}^T\mathrm{C}_i\Theta_{k,i}. \tag{19}$$

For this inequality to be satisfied, each node $l \in \mathcal{N}_{k,i}^{+}$ must satisfy the condition $\|\mathrm{C}_i\xi_{lk,i}\|^2 \le -2\Delta\widehat{w}_{k,i}{}^T\mathrm{C}_i\xi_{lk,i}$. We relax the condition by taking the average of $\|\xi_{lk,i}\|^2$ with respect to all M elements. We then obtain an element-wise difference

$$\left|\sum_{l\in\mathcal{N}_{k,i}^{+}} c_{lk,i}\sqrt{\frac{\|\xi_{lk,i}\|^2}{M}}\right| \le |\Delta\widehat{w}_{k,i}{}^{(m)}|, \tag{20}$$

which holds for each element $\Delta\widehat{w}_{k,i}{}^{(m)} \in \Delta\widehat{w}_{k,i}$. Recalling that $\sqrt{\|\xi_{lk,i}\|^2/M} \le \sqrt{\gamma_{k,i}}$ and $\sum_{l\in\mathcal{N}_{k,i}^{+}} c_{lk,i} = 1$, we have $\left|\sum_{l\in\mathcal{N}_{k,i}^{+}} c_{lk,i}\sqrt{\|\xi_{lk,i}\|^2/M}\right| \le \sqrt{\gamma_{k,i}}$, and if $\gamma_{k,i} \le 4(\Delta\widehat{w}_{k,i}{}^{(m)})^2$ for each element $\Delta\widehat{w}_{k,i}{}^{(m)} \in \Delta\widehat{w}_{k,i}$, then the transient squared $\|\Delta w_{k,i}\|^2$ of the d-LMS system will be smaller than $\|\Delta\widehat{w}_{k,i}\|^2$ of a reliable approximate optimal vector. This is equivalent to $\gamma_{k,i} \le 4(\Delta\widehat{w}_{k,i}{}^{(\min)})^2$, where $\Delta\widehat{w}_{k,i}{}^{(\min)}$ is the minimum element of $\Delta\widehat{w}_{k,i}$.

Improper values will lead to inaccurate clustering and poor estimation performance due to less information being available from cooperative nodes. However, in practical applications, the actual value of w_k^o, namely, $(\Delta\widehat{w}_{k,i}{}^{(m)})^2$, is unknown and should be estimated via the sample variance. However, beyond that, the minimum element $\Delta\widehat{w}_{k,i}{}^{(\min)} \in \Delta\widehat{w}_{k,i}$ will change in real time. We relax the condition without loss of accuracy by taking as an approximation an average of the relatively reliable approximate optimal vector, which is the approximate variance of the local optimum $\widehat{w}_{k,i}^{o}$ with respect to

TABLE I. Algorithm 1: Distributed estimation with adaptive clustering based on element-wise distance over multitask networks.

Each agent $k \in S$ follows these steps:
Initialize:
Start with $w_{k,-1} = 0$.
For $i = 1{:}T$
(1) Adaptation:

$$\varphi_{k,i} = w_{k,i-1} + \mu_k u_{k,i}^T (d_k(i) - u_{k,i}w_{k,i-1})$$

(2) Clustering detection and updating $\mathcal{N}_{k,i}^{+}$ according to Algorithm 2 (Table II)
(3) Updating combination weight $c_{lk,i}$ according to (23)
(4) Combination:

$$w_{k,i} = \sum_{l\in\mathcal{N}_{k,i}^{+}} c_{lk,i}\varphi_{l,i}$$

end

TABLE II. Algorithm 2: Distributed adaptive clustering based on element-wise distance.

Each agent $k \in S$ follows these steps:
Initialize:
Start with $\widehat{w}^o_{k,-1} = \hat{\varsigma}_{k,-1} = 0$, and $\hat{\delta}^2_{k,-1} = 0$.
For $i = 1{:}T$

(1) Making a one-step approximation to obtain an approximate optimal value $\widehat{w}^o_{k,i}$:

$$\widehat{w}^o_{k,i} = \varphi_{k,i} - \mu_k \nabla_k J_k(w)\big|_{w=\varphi_{k,i}} \approx \varphi_{k,i} + \mu_k u^T_{k,i}(d_k(i) - u_{k,i}\varphi_{k,i})$$

with $\varphi_{k,i}$ updating in (6)

(2) Updating adaptive clustering threshold $\gamma_{k,i}$:

$$\hat{\varsigma}_{k,i} = (1-a)\hat{\varsigma}_{k,i-1} + a\widehat{w}^o_{k,i-1}$$
$$\hat{\delta}^2_{k,i} = (1-a)\hat{\delta}^2_{k,i-1} + a(\|\widehat{w}^o_{k,i-1} - \hat{\varsigma}_{k,i-1}\|^2/M)$$
$$\gamma_{k,i} = 4\hat{\delta}^2_{k,i}$$

(3) Clustering detection with element-wise distance $\theta_{lk,i}$ and updating $\mathcal{N}^+_{k,i}$: *for* $\forall\, l \in \mathcal{N}_k$

$$T_{lk,i} = \theta_{lk,i} \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\lessgtr}} \lambda_k \gamma_{k,i}$$
$$\mathcal{N}^+_{k,i} \triangleq \{l \in \mathcal{N}_k \mid T_{lk,i} = \theta_{lk,i} \le \lambda_k \gamma_{k,i}\}$$

with $\theta_{lk,i} = \|\widehat{w}^o_{k,i} - \varphi_{l,i}\|^2/M$
end
end

all the M elements:

$$\hat{\varsigma}_{k,i} = (1-a)\hat{\varsigma}_{k,i-1} + a\widehat{w}^o_{k,i},$$
$$\hat{\delta}^2_{k,i} = (1-a)\hat{\delta}^2_{k,i-1} + a(\|\widehat{w}^o_{k,i} - \hat{\varsigma}_{k,i}\|^2/M). \tag{21}$$

where $\hat{\varsigma}_{k,i}$ and $\hat{\delta}^2_{k,i}$ are the mean and the averaged element-wise sample variance of $\widehat{w}^o_{k,i}$, respectively. Each agent takes advantage of $\widehat{w}^o_{k,i}$ at time i to maintain a sample dynamics standard mean as the running average of its local approximation, and also estimates the average element-wise sample variance by means of smoothing recursions as a real-time clustering threshold.

Based on the approximate optimal vector $\widehat{w}^o_{k,i}$ obtained by the one-step approximation in Sec. IV A, the specific form of adaptive clustering threshold for reference as well as the relationship between the real-time clustering threshold and the clustering detection statistics are derived.

Obviously, from (21), these two clustering reference values obtained by the exponential smoothing method will change in real time. We can see that through the distributed diffusion strategy, each node applies an adaptive gain to the measurement during the estimation update process, and the impact of abnormal and damaged measurements will give rise to fluctuations in the adaptive gain, and thus also in the intermediate estimate as well as the instantaneous estimate. That is to say, in a nonstationary multitask environment, task anomalies, damaged measurements, or time-varying tasks can cause fluctuations in the adaptive clustering threshold. When the environment is highly nonstationary, exponential smoothing is a good choice to keep the algorithm alive by assigning a reasonable a. In this way, we can extend the clustering framework to an unstable multitask environment.

Finally, we set

$$\gamma_{k,i} = 4\hat{\delta}^2_{k,i}. \tag{22}$$

The sample statistics converge to unbiased estimates of their corresponding true values, since $\widehat{w}^o_{k,i}$ approaches w^o_k as $i \to \infty$. Hence, the setting for $\gamma_{k,i}$ will ensure the accuracy of clustering in real time.

C. Clustering detection with element-wise statistics

The averaged element-wise distance $\theta_{lk,i}$ between node l's intermediate estimate $\varphi_{l,i}$ and a relatively reliable approximate optimal vector of the local optimum, $\widehat{w}^o_{k,i}$, should be calculated as a clustering detection statistic for accurate clustering at time i as

$$\theta_{lk,i} = \|\widehat{w}^o_{k,i} - \varphi_{l,i}\|^2 M \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\lessgtr}} \gamma_{k,i} \quad \text{for } l \in \mathcal{N}_k, \tag{23}$$

where $\mathcal{H}_0$ and $\mathcal{H}_1$ denote the hypothesis that node l belongs to the subneighbors that are in the same cluster as node k and the opposite hypothesis, respectively. It can be seen from the above analysis that clustering has been restricted to a relatively strict range, so that it may lead to a noncooperation situation, which makes the gain not worth the loss of estimation accuracy. Consequently, we add a relaxation factor $\lambda_k > 1$ to relax clustering restrictions by making a tradeoff between maximum cooperation for steady-state misalignment of estimation and clustering accuracy. Thus, the

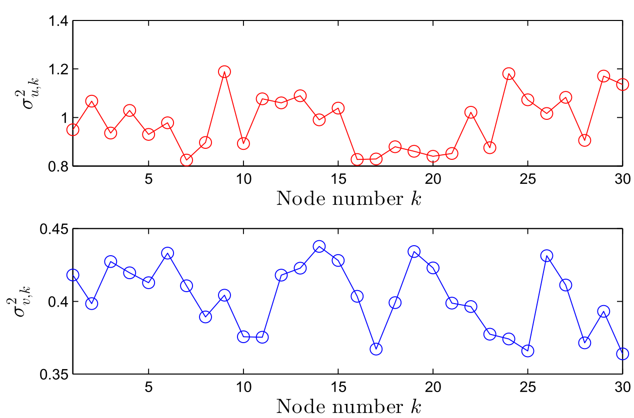

FIG. 2. Simulation profiles across all agents in the network.

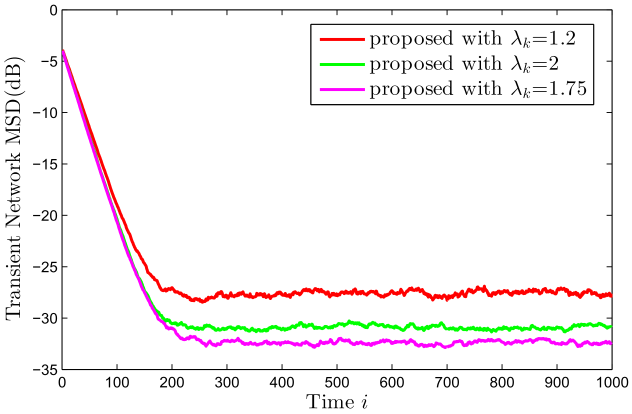

FIG. 3. Performance of distributed estimation. Transient-state average MSDs for recursions with various λ_k.

corresponding clustering hypothesis is derived as

$$T_{lk,i} = \theta_{lk,i} \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\lessgtr}} \lambda_k \gamma_{k,i}. \tag{24}$$

Clustering detection at time i will lead to the following set of real-time subneighbors that are in the same cluster as node k and able to cooperate:

$$\mathcal{N}_{k,i}^{+} \triangleq \{l \in \mathcal{N}_{k,i}^{+} \mid T_{lk,i} = \theta_{lk,i} \leq \lambda_k \gamma_{k,i}\}. \tag{25}$$

Recalling (22), (25) suggests that the estimator $\varphi_{l,i}$, for $\forall l \in \mathcal{N}_{k,i}^{+}$ should be close to $\widehat{w}_{k,i}^{o}$ to within an acceptable degree.

Previous algorithms using the Euclidean distance between $\widehat{w}_{k,i}^{o}$ and $\varphi_{l,i}$ to construct combination weights will not be able to distinguish clusters from each other if the cluster objectives $\{w_k^o\}$ are not significantly distinct, thereby resulting in poor estimation performance. By contrast, in our clustering method, because the adaptive clustering threshold ensures accuracy, diversity of cluster vectors is not a strict requirement.

We use the Metropolis rule as a combination rule in which the combination weight $c_{l,k,i}$ at time i is defined as

$$c_{lk,i} = \begin{cases} \dfrac{1}{\max(n_{k,i}, n_{l,i})} & \text{if } l \in \mathcal{N}_{k,i}^{+} \backslash \{k\}, \\ 1 - \displaystyle\sum_{l \in \mathcal{N}_{k,i}^{+} \backslash \{k\}} c_{lk,i} & \text{if } l = k, \\ 0 & \text{otherwise.} \end{cases} \tag{26}$$

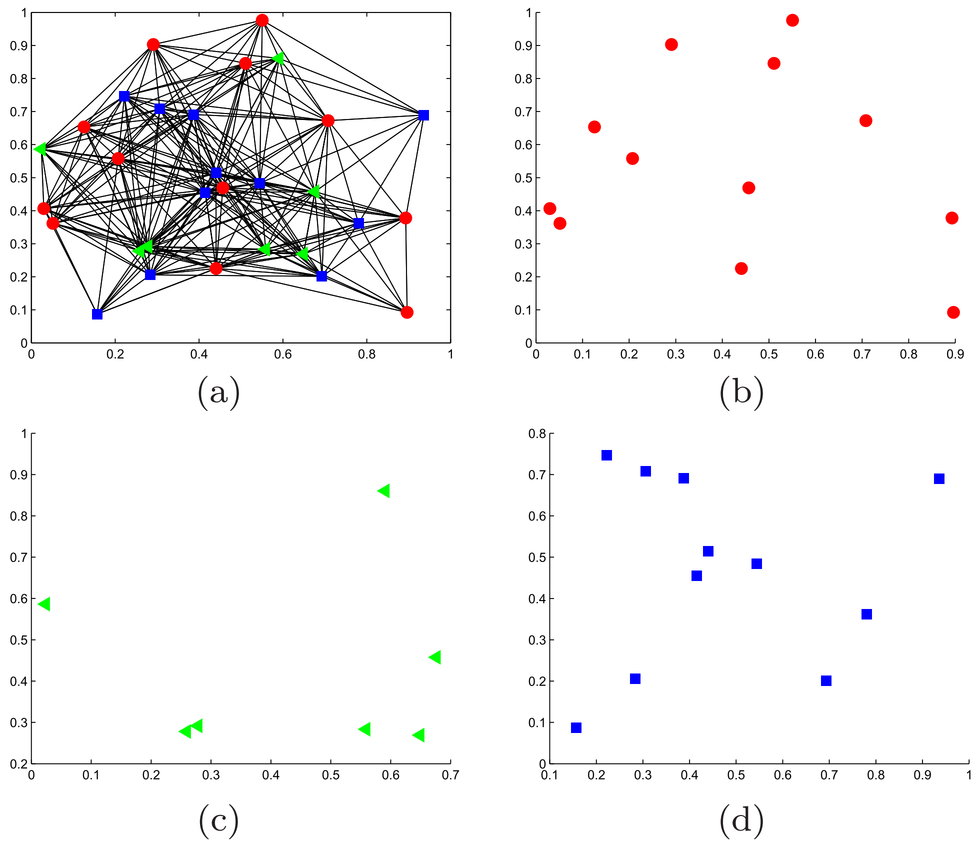

FIG. 4. Network topology in the initial stage with no prior cluster information in case 1: (a) initial global network topology; (b), (c), and (d) initial topologies of clusters 1, 2, and 3, respectively.

Our distributed estimation algorithm using the adaptive clustering method based on element-wise distance over multitask networks and our distributed adaptive clustering algorithm based on element-wise distance are summarized in pseudocode in Tables I and II.

V. SIMULATION RESULTS

In this section, simulations and a performance analysis of adaptive clustering and distributed estimation are provided. We simulate over a connected network, which is a randomly generated topology with $N = 30$ nodes. In the connected network that we consider, there is no isolated node without neighbors, and each node of this network has at least five single-step accessible neighbors as well as at least one neighbor belonging to the same cluster as it does. The regression inputs $u_{k,i}$ with length $L = 2$ are randomly generated by a Gaussian distribution with zero mean and covariance matrix $R_{u,k} = \sigma_{u,k}^2 I_L$. The model noises $v_k(i)$ with covariance matrix $R_{v,k} = \sigma_{v,k}^2 I_L$ are also i.i.d. zero-mean Gaussian random variables. The variances of $u_{k,i}$ and $v_k(i)$ are plotted in Figs. 2(a) and 2(b), respectively.

Individual comparisons of our proposed clustering algorithm for $\lambda_k = 1.2, 1.75, 2$, and $\mu_k = 0.01$, and $a = 0.4$ are also provided in Fig. 3. It can be concluded that the performance of the algorithm is optimal when $\lambda_k = 1.75$. The clustered multitask adaptive estimation strategy[14] and the multitask adaptive estimation without clustering strategy[16] are taken as examples to compare with the proposed multitask adaptive estimation with adaptive clustering strategy for different step sizes $\mu_k = 0.01, 0.02, 0.03$ and $\lambda_k = 1.75$ in both stationary and nonstationary environments. The simulations are divided into two cases according to the significance of the differences among the cluster optima $w^o \triangleq \mathrm{col}\{w_k^o\}_{k=1}^N$.

A. Stationary environment

Case 1: No significant differences in the setting of cluster optima. The time-invariant cluster optima are set as follows:

$$w_k^o = \begin{cases} [0.54, -0.48]^T, & k = 1, \ldots, 12 \ \text{(cluster 1)}, \\ [0.47, -0.5]^T, & k = 13, \ldots, 19 \ \text{(cluster 2)}, \\ [0.56, -0.4]^T, & k = 20, \ldots, 30 \ \text{(cluster 3)}. \end{cases}$$

Here, 30 nodes are divided into three clusters corresponding to the colors red, green, and blue, respectively. Figure 4 shows the network topology in the initial stage. All possible links are shown in Fig. 4(a), and as depicted in Figs. 4(b)–4(d), nodes belonging to different clusters do not form cluster structures in the initial stage, because it is not yet clear whether subneighbors are in the same cluster or in different clusters of nodes, and agents have no cluster information

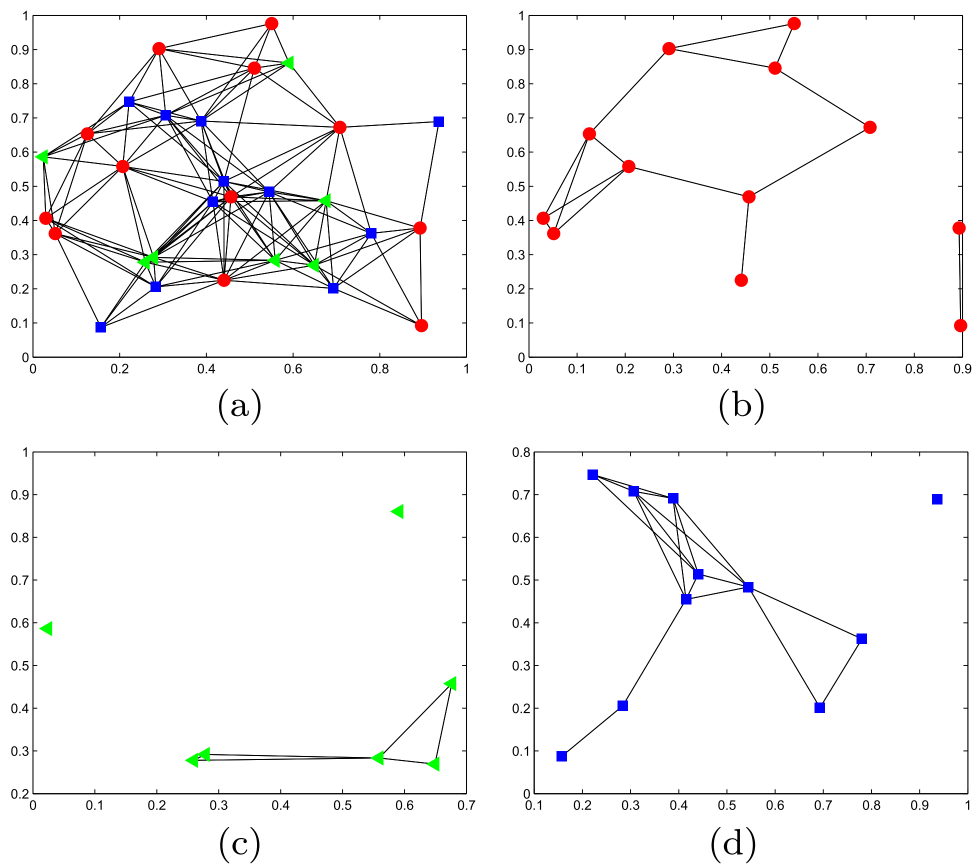

FIG. 5. Performance of adaptive clustering. Network topology in a steady state after the clustering process in case 1: (a) global network topology in steady state; (b), (c), and (d) topologies of clusters 1, 2, and 3, respectively, in a steady state.

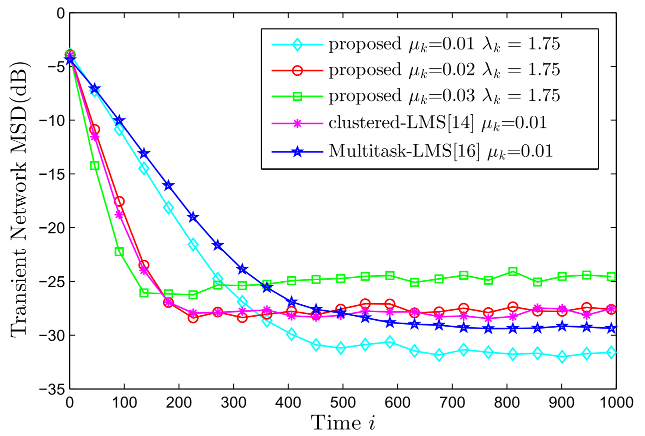

FIG. 6. Performance of distributed estimation. Network transient-state average MSDs for recursions in case 1.

at all in advance. The clustering statistics $\|\widehat{w}^o_{k,i} - \varphi_{l,i}\|^2/M$ are time-variant in the initial clustering phase, but as $i \to \infty$, the clustering statistics will become time-invariant, so the results of clustering detection will be steady. The steady-state network topology with clustered structure and subnetworks for each cluster after the clustering process, with all the links between subneighbors in different clusters broken up, are depicted in Fig. 5.

We then compare our proposed multitask adaptive estimation with adaptive clustering strategy with the other two strategies mentioned above. The network transient-state average MSDs are depicted in Fig. 6. It is observed that the clustered estimation strategy[14] is obviously not suitable for an unknown cluster multitask environment, and it gives the worst MSD performance. In case 1, there are no significant differences in the setting of cluster optima. Thus, our proposed estimation strategy with adaptive clustering is only slightly better than the strategy in Ref. 16, with an MSD of almost 3 dB.

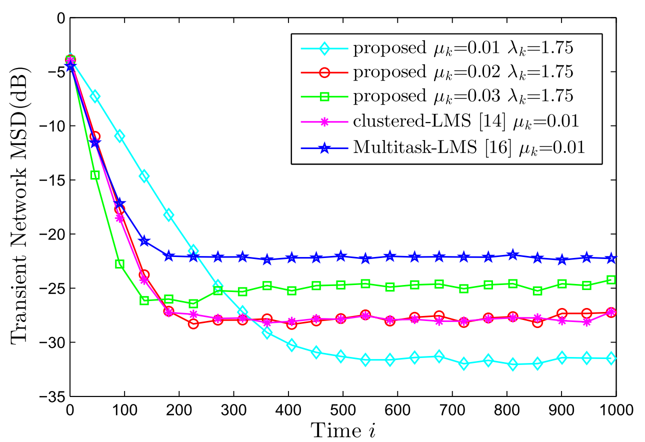

FIG. 7. Performance of distributed estimation. Network transient-state average MSDs for recursions in case 2.

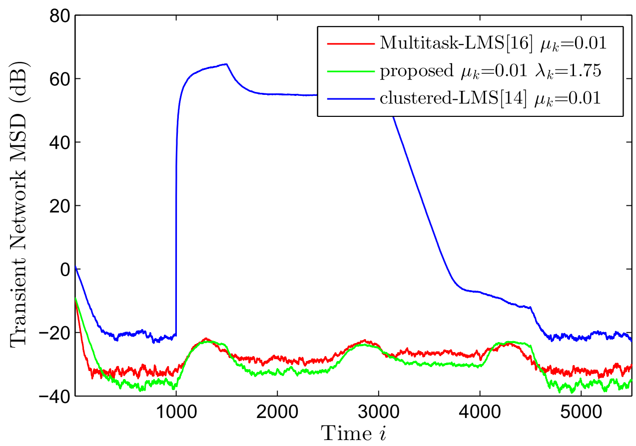

FIG. 8. Performance of distributed estimation. Comparison of network MSD behavior in a nonstationary multitask environment.

Case 2: Significant differences in the setting of cluster optima. In this situation, the corresponding cluster optima are set as follows:

$$w_k^o = \begin{cases} [0.5, 0.4]^T, & k = 1, \ldots, 6 \ \text{(cluster 1)}, \\ [-0.1, -0.2]^T, & k = 7, \ldots, 12 \ \text{(cluster 2)}, \\ [0.3, -0.3]^T, & k = 13, \ldots, 21 \ \text{(cluster 3)}, \\ [-0.8, 0.5]^T, & k = 22, \ldots, 30 \ \text{(cluster 4)}. \end{cases}$$

The nodes are divided into four clusters. Note that the distance in each dimension between any two local optima is more obvious than in case 1.

From the cluster structure evolution in Figs. 4 and 5 and the network MSD shown in Figs. 6 and 7, we find that whether the differences among local optima of each cluster $\{w_k^o\}$ are significant has little impact on the performance of our proposed clustering strategy and the distributed estimation algorithm with adaptive clustering method. However, it has a very strong effect in the case of the algorithm in Ref. 16 that uses an adaptive combination weight for clustering. It introduces a large bias into the estimation when diffusion LMS is used over the whole multitask network according to the adaptive combination weight among all of the neighbors, because the distance between any two of the $\{w_k^o\}$ is large and results in the introduction of multiple task-irrelevant neighbors. As a consequence, when the local optima $\{w_k^o\}$ of each cluster are significantly different from each other, the algorithm in Ref. 16 has a poor MSD that is almost 5 dB worse than in case 1. By contrast, with our proposed adaptive clustering method, there are few constraints on the diversity of local optima of each cluster. We find from Figs. 4 and 5 that a better clustering result is obtained and from Figs. 6 and 7 that fluctuations in the MSD performance are reduced.

B. Nonstationary environment

Consider a time-variant multitask environment where the clusters vary over time. Similarly to the setting in Ref. 16, from instant 1 to instant 1000, the network consists of a single cluster with a unique optimum to estimate. From instant 1001 to instant 4000, the nodes

are split into two and then into four clusters. Finally, from instant 4001, the nodes aggregate into a single cluster with another unique local optimum to estimate. For example, the first element of $\{w_k^o\}$ for the time interval from 1 to 2500 can be expressed as

$$w_{k,1}^o(i) = \begin{cases} 0.3, & i = 1, \ldots, 1000, \\ 0.3 + \frac{0.5-0.3}{500}(i - 1000), & i = 1001, \ldots, 1500, \\ 0.5, & i = 1501, \ldots, 2500. \end{cases}$$

As can be seen from Fig. 8, our proposed clustering strategy has a better estimation performance than the algorithm in Ref. 16, and the clustered multitask algorithm in Ref. 14 is completely unadapted to a nonstationary environment.

VI. CONCLUSION

In this work, we have focused on distributed adaptive estimation through adaptive clustering over multitask networks. We have devised an adaptive clustering method that is constructed through a real-time clustering hypothesis test based on an adaptive clustering threshold as reference using the averaged element-wise distance for accurate clustering, instead of imposing a static quantitative clustering threshold. We have proposed a distributed estimation algorithm using the adaptive clustering method. We have presented simulation results illustrating the superior performance of our proposed algorithms, which represent a more suitable, robust, and accurate approach than previous strategies, regardless of differences among local optima in both stationary and nonstationary environments.

ACKNOWLEDGMENTS

This work was supported in part by the National Key R&D Program of China (No. 2018YFB1306600), and the Fundamental Research Funds for the Central Universities (Nos. XDJK2020B034 and SWU019029).

REFERENCES

[1] S. Kar and J. M. F. Moura, "Distributed consensus algorithms in sensor networks: Link failures and channel noise," IEEE Trans. Signal Process **57**(1), 355–369 (2009).
[2] P. Braca, S. Marano, V. Matta, and P. Willett, "Asymptotic optimality of running consensus in testing binary hypotheses," IEEE Trans. Signal Process. **58**(2), 814–825 (2010).
[3] D. P. Bertsekas, "A new class of incremental gradient methods for least squares problems," SIAM J. Optim. **7**(4), 913–926 (1997).
[4] A. Nedic and D. P. Bertsekas, "Incremental subgradient methods for nondifferentiable optimization," SIAM J. Optim. **12**(1), 109–138 (2001).
[5] C. G. Lopes and A. H. Sayed, "Diffusion least-mean squares over adaptive networks: Formulation and performance analysis," IEEE Trans. Signal Process. **56**(7), 3122–3136 (2008).
[6] F. S. Cattivelli and A. H. Sayed, "Diffusion LMS strategies for distributed estimation," IEEE Trans. Signal Process. **58**(3), 1035–1048 (2010).
[7] S.-Y. Tu and A. H. Sayed, "Diffusion strategies outperform consensus strategies for distributed estimation over adaptive netowrks," IEEE Trans. Signal Process. **60**(12), 6217–6234 (2012).
[8] A. H. Sayed, "Adaptive networks," Proc. IEEE **102**(4), 460–497 (2014).
[9] A. H. Sayed, "Adaptation, learning, and optimization over networks," Found. Trends® Mach. Learn. **7**(4C5), 311–801 (2014).
[10] A. G. Dimakis, S. Kar, J. M. F. Moura, M. G. Rabbat, and A. Scaglione, "Gossip algorithms for distributed signal processing," Proc. IEEE **98**(11), 1847–1864 (2010).
[11] R. Olfati-Saber, J. A. Fax, and R. M. Murray, "Consensus and cooperation in networked multi-agent systems," Proc. IEEE **95**(1), 215–233 (2007).
[12] J. Chen and A. H. Sayed, "Diffusion adaptation strategies for distributed optimization and learning over networks," IEEE Trans. Signal Process. **60**(8), 4289–4305 (2012).
[13] X. Zhao and A. H. Sayed, "Clustering via diffusion adaptation over networks," in *2012 3rd International Workshop on Cognitive Information Processing (CIP)* (IEEE, 2012), pp. 1–6.
[14] J. Chen, C. Richard, and A. H. Sayed, "Multitask diffusion adaptation over networks," IEEE Trans. Signal Process. **62**(16), 4129–4144 (2014).
[15] W. Huang, X. Yang, and G. Shen, "Communication-reducing diffusion lms algorithm over multitask networks," Inf. Sci. **382**, 115–134 (2017).
[16] J. Chen, C. Richard, and A. H. Sayed, "Diffusion LMS over multitask networks," IEEE Trans. Signal Process. **63**(11), 2733–2748 (2015).
[17] L. Dang, B. Chen, S. Wang *et al.* "Kernel Kalman filtering with conditional embedding and maximum correntropy criterion," IEEE Trans. Circuits Syst. I **66**(11), 4265–4277 (2019).
[18] B. Chen, L. Dang, Y. Gu *et al.*, "Minimum error entropy Kalman filter," arXiv:1904.06617, 2019.
[19] F. Chen and X. Shao, "Broken-motifs diffusion LMS algorithm for reducing communication load," Signal Process. **133**, 213–218 (2017).
[20] F. Chen, X. Li, S. Duan, L. Wang, and J. Wu, "Diffusion generalized maximum correntropy criterion algorithm for distributed estimation over multitask network," Digit. Signal Process. **81**, 16–25 (2018).
[21] X. Li, Q. Shi, S. Xiao *et al.*, "A robust diffusion minimum kernel risk-sensitive loss algorithm over multitask sensor networks," Sensors **19**(10), 2339 (2019).
[22] M. Wan, L. Li, J. Xiao *et al.*, "Data clustering using bacterial foraging optimization," J. Intell. Inf. Syst. **38**(2), 321–341 (2012).
[23] M. Wan, C. Wang, L. Li *et al.*, "Chaotic ant swarm approach for data clustering," Appl. Soft Comput. **12**(8), 2387–2393 (2012).
[24] S.-Y. Tu and A. H. Sayed, "Distributed decision-making over adaptive networks," IEEE Trans. Signal Process. **62**(5), 1054–1069 (2014).
[25] J. Chen and A. H. Sayed, "Distributed Pareto optimization via diffusion strategies," IEEE J. Sel. Top. Signal Process. **7**(2), 205–220 (2013).
[26] X. Wu, Y. Tang, J. Cao, and X. Mao, "Stability analysis for continuous-time switched systems with stochastic switching signals," IEEE Trans. Automat. Contr. **63**(9), 3083–3090 (2018).
[27] Y. Tang, X. Wu, P. Shi, and F. Qian, "Input-to state stability for nonlinear systems with stochastic impulses," Automatica **113**(3), 108766 (2020).
[28] Y. Liu and C. Li, "Secure distributed estimation over wireless sensor networks under attacks," IEEE Trans. Aerosp. Electron. Syst. **54**(4), 1815–1831 (2018).