Improving the Performance of Multitask Diffusion APA via Controlled Inter-Cluster Cooperation

Vinay Chakravarthi Gogineni and Mrityunjoy Chakraborty, *Senior Member, IEEE*

***Abstract*— In this paper, we consider the problem of estimating multiple parameter vectors over a sensor network in a multitasking framework and under temporally-correlated input conditions. For this, an efficient clustered multitask diffusion affine projection algorithm (APA) is proposed that enjoys both intra-cluster and inter-cluster cooperation via diffusion. It is, however, shown that while collaboration in principle is a useful step to enhance the performance of a network, uncontrolled mode of inter-cluster collaboration can at times be detrimental to its convergence performance, especially near steady-state. To overcome this, a controlled form of inter-cluster collaboration is proposed by means of a control variable which helps in maintaining the collaboration in right direction. The proposed controlled multitask strategy attains improved performance in terms of both transient and steady-state mean square deviation (MSD) vis-a-vis existing algorithms, as also confirmed by simulation studies. We carry out a detailed performance analysis of the proposed algorithm, obtain stability bounds for its convergence in both mean and mean-square senses, and derive expressions for the network level MSD. Simulation results reveal that the proposed scheme performs consistently well even in the absence of cluster information.**

***Index Terms*— Multitask network, distributed adaptive estimation, block maximum norm, adaptive diffusion networks, affine projection algorithm.**

I. Introduction

ADAPTIVE networks consist of interconnected nodes that adaptively estimate certain parameter vector(s) of interest, by deploying some collaboration among the neighboring nodes [1]–[3]. In a single-task network, all nodes collectively estimate a single optimal parameter vector (i.e., each node is engaged in a common task, e.g., point target localization [1]), for which, several useful modes of collaboration have been proposed and analyzed recently, like incremental [4], [5], consensus [6], [7] and diffusion strategies [8]–[11]. Of these, diffusion strategies are simple but more efficient as compared to the other two for distributed adaptive estimation [12].

Beside single-task scenarios, in some applications, adaptive networks are required to estimate multiple parameter vectors simultaneously and are called multitask networks. For example, in distributed active noise control applications [13], agents need to determine different but related active noise control filters. Similarly, in node-specific cooperative spectrum sensing [14] and node-specific speech enhancement [15], multiple parameter vectors need to be estimated jointly in collaborative fashion. In a multitask network, the nodes are grouped into clusters and nodes within the same cluster estimate a common parameter vector [16]. Different clusters generally carry out different (though related) tasks and the relationship between these tasks is unknown. The estimation still needs to be carried out cooperatively across the network because the data across the clusters may be correlated and therefore, cooperation across clusters can be beneficial. In other words, a multitask network employs both inter-cluster and intra-cluster cooperation. In [17], [18], a least mean square (LMS) based multitask diffusion algorithm has been presented, which is studied in [19] in presence of random link failures and changing topology, and in [20], a robust learning approach for the same is presented. Separately, using the robustness of the affine projection algorithm (APA) [21] against colored input, an APA based diffusion multitask estimation scheme has been proposed in [22]. In the aforementioned works, it is shown that the network level mean square deviation (MSD) performance of multitask diffusion schemes depends on the extent of inter-cluster cooperation present which is controlled by certain regularization strength parameter and regularization coefficients between the inter-cluster nodes. In [22], it is also observed that certain choices of these weights may lead to poor performance compared to the non-cooperation case.

In this paper, we propose a multitask diffusion APA with efficient inter-cluster cooperation. Our main contributions here are as follows:

1) It is shown that while collaboration in principle is a useful step to enhance network performance, uncontrolled mode of inter-cluster collaboration can, however, lead to deterioration of network convergence behavior, especially near steady-state. To overcome this, we propose an improved clustered multitask diffusion strategy that uses a control variable to maintain cooperation among neighboring clusters in the right direction. The proposed algorithm exhibits faster convergence rate and lesser steady-state MSD than the state-of-the-art.

Manuscript received July 12, 2019; revised October 25, 2019 and November 14, 2019; accepted November 19, 2019. Date of publication December 30, 2019; date of current version March 4, 2020. This work was supported in part by the Science and Engineering Research Board (SERB), Government of India, and in part by the European Research Consortium for Informatics and Mathematics (ERCIM) Alain Bensoussan Fellowship Programme, EU. This article was recommended by Associate Editor D. Comminiello. *(Corresponding author: Mrityunjoy Chakraborty.)*

V. C. Gogineni was with the Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology Kharagpur, Kharagpur 721302, India. He is now with the Department of Electronic Systems, Norwegian University of Science and Technology (NTNU), 7491 Trondheim, Norway (e-mail: vinaychakravarthi@ece.iitkgp.ernet.in).

M. Chakraborty is with the Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology Kharagpur, Kharagpur 721302, India (e-mail: mrityun@ece.iitkgp.ernet.in).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSI.2019.2957342

2) Extending the energy conservation approach based APA analysis [23] to the distributed implementation, we carry out a detailed performance analysis of the proposed improved clustered multitask diffusion APA, where we derive stability bounds for its convergence in both mean and mean-square senses, and derive expressions for the network level instantaneous and steady-state MSD that explain the effects of various design parameters on the network performance.
3) We demonstrate the effectiveness of the proposed algorithm through detailed simulations in a system identification context.

The rest of the paper is organized as follows. In Section II, we introduce the network model along with some prior related works, and present the proposed improved multitask diffusion strategy. Next, in Section III, we carry out a detailed performance analysis of the proposed algorithm and obtain stability bounds for its convergence in both mean and mean-square senses. Detailed simulation results in support of the proposed algorithm are provided in Section IV. Section V concludes the paper.

II. Network Model and Proposed Algorithm

A. Clustered Multitask Diffusion Affine Projection Algorithm

We consider here a clustered multitask network with N nodes that are grouped into Q clusters. Each node k has access to the input signal $u_k(n)$ and the observable output $d_k(n)$ that are assumed to be related via a linear model

$$d_k(n) = \mathbf{u}_k^T(n)\mathbf{w}_k^\star + \vartheta_k(n), \tag{1}$$

where $\mathbf{w}_k^*$ is the $L\times 1$ optimal parameter vector to be estimated at node k, $\mathbf{u}_k(n) = [u_k(n), u_k(n-1), \cdots, u_k(n-L+1)]^T$ is the input data vector and $\vartheta_k(n)$ is an observation noise with zero mean and variance $\sigma^2_{\vartheta,k}$ which is taken to be temporally and spatially i.i.d. and independent of input $\mathbf{u}_l(m)$ for all n, m and k, l. The nodes that are grouped in the same cluster C_q, $q = 1, 2, \cdots, Q$, estimate the same $L \times 1$ filter coefficient vector $\mathbf{w}^*_{C_q}$, implying $\mathbf{w}_k^* = \mathbf{w}^*_{C_q}$ for $k \in C_q$. Further, the optimal parameter vectors (also termed as tasks) of neighboring clusters $\mathbf{w}^\star_{C_q}$ and $\mathbf{w}^\star_{C_r}$ are different but are related, implying $\mathbf{w}^\star_{C_q} \sim \mathbf{w}^\star_{C_r}$ if clusters C_q and C_r are connected. We use the notation $C(k)$ to denote the cluster to which node k belongs, $C(k) \in \{C_1, C_2, \cdots, C_Q\}$.

Each node k estimates the corresponding $\mathbf{w}_k^*$ by updating a weight vector $\mathbf{w}_k(n)$ following the adapt-then-combine (ATC) based clustered multitask diffusion APA [22], which follows directly from [17] and is given as follows:

Adaptation:

$$\boldsymbol{\psi}'_k(n+1) = \mathbf{w}_k(n) + \mu\mathbf{U}_k(n)\big(\epsilon\mathbf{I}_P + \mathbf{U}_k^T(n)\mathbf{U}_k(n)\big)^{-1}\mathbf{e}_k(n),$$

Combination (Inter-Cluster):

$$\boldsymbol{\psi}_k(n+1) = \boldsymbol{\psi}'_k(n+1) + \mu\eta \sum_{l\in\mathcal{N}_k\setminus C(k)} \rho_{kl}(\mathbf{w}_l(n) - \mathbf{w}_k(n)),$$

Combination (Intra-Cluster):

$$\mathbf{w}_k(n+1) = \sum_{l\in\mathcal{N}_k\cap C(k)} a_{lk}\boldsymbol{\psi}_l(n+1), \tag{2}$$

where $\mathbf{U}_k(n) = [\mathbf{u}_k(n), \mathbf{u}_k(n-1), \cdots, \mathbf{u}_k(n-P+1)]$ is the input signal matrix, $\mathbf{e}_k(n) = \mathbf{d}_k(n) - \mathbf{U}_k^T(n)\mathbf{w}_k(n) \equiv [e_k(n), e_k(n-1), \cdots, e_k(n-P+1)]^T$ is the error vector, $\mathbf{d}_k(n) = \mathbf{U}_k^T(n)\mathbf{w}_k^* + \boldsymbol{\vartheta}_k(n) \equiv [d_k(n), d_k(n-1), \cdots, d_k(n-P+1)]^T$ is the desired response vector, and $\boldsymbol{\vartheta}_k(n) = [\vartheta_k(n), \vartheta_k(n-1), \cdots, \vartheta_k(n-P+1)]^T$ is the observation noise vector, all at the k^{th} node, $k = 1, 2, \cdots, N$, with P being the projection order. The symbol ϵ is a small positive constant deployed to avoid inversion of a rank deficient matrix $\mathbf{U}_k^T(n)\mathbf{U}_k(n)$. The symbol $\mathcal{N}_k$ denotes the neighborhood of node k including k. The small positive constant η is a regularization strength parameter and the non-negative coefficients ρ_{kl} adjust the regularizer strength between inter-cluster nodes k and l. The non-negative coefficients ρ_{kl} are chosen to satisfy the following conditions [17]:

$$\sum_{l=1}^{N}\rho_{kl} = 1, \quad \text{and} \quad \begin{cases}\rho_{kl} > 0, & \text{if } l \in \mathcal{N}_k \setminus C(k),\\ \rho_{kl} = 0, & \text{if } l \notin \mathcal{N}_k \setminus C(k).\end{cases} \tag{3}$$

Further, $\rho_{kk} = 1$ if $\mathcal{N}_k \setminus C(k) = \emptyset$. The combination coefficients a_{lk} are non-negative and are given by

$$\sum_{l=1}^{N} a_{lk} = 1, \quad \text{and} \quad \begin{cases}a_{lk} > 0, & \text{if } l \in \mathcal{N}_k \cap C(k),\\ a_{lk} = 0, & \text{otherwise.}\end{cases} \tag{4}$$

[A matrix with its elements a_{lk} satisfying (4) is called *left stochastic*; also, a matrix is *right stochastic* if its transpose is left stochastic.] Several methods exist in literature to select the coefficients a_{lk}, e.g., averaging rule, Metropolis rule etc. [2].

B. Improved Clustered Multitask Diffusion Affine Projection Algorithm

The clustered multitask diffusion strategy presented in (2), however, induces two problems:

1) Assume, at time index n, node $l \in \mathcal{N}_k \setminus C(k)$ exhibits poor performance over node k. The clustered multitask diffusion scheme does not take this into account and aimlessly allows node k to learn from poorly performing node l. This affects the transient performance of the algorithm (2).
2) For all $l \in \mathcal{N}_k \setminus C(k)$, we have $\mathbf{w}_l^* \sim \mathbf{w}_k^*$, i.e., the neighboring cluster tasks are only having some kind of similarity relationship, but are not exactly the same. This means, ideally, the inter-cluster cooperation has to be terminated near convergence. However, the clustered multitask diffusion strategy (2) continues the inter-cluster cooperation between k and l even in the steady-state, and the cooperation is in proportion to the value of η. This may hamper the steady-state performance of the algorithm.

To address these problems, motivated from [24], we introduce a control variable $\delta_{kl}(n)$ to regulate the learning rate during inter-cluster cooperation, given as,

$$\delta_{kl}(n) = \frac{1}{2}\big(1 + sgn\big(\sigma_k^2(n) - \sigma_{kl}^2(n)\big)\big), \quad l \in \mathcal{N}_k \setminus C(k), \tag{5}$$

where $sgn(\cdot)$ is the well known signum function. The error variances $\sigma_k^2(n)$ and $\sigma_{kl}^2(n)$ are recursively updated as

$$\sigma_k^2(n) = \gamma\,\sigma_k^2(n-1) + (1-\gamma)\left(d_k(n) - \mathbf{u}_k^T(n)\mathbf{w}_k(n)\right)^2,$$
$$\sigma_{kl}^2(n) = \gamma\,\sigma_{kl}^2(n-1) + (1-\gamma)\left(d_k(n) - \mathbf{u}_k^T(n)\mathbf{w}_l(n)\right)^2 \quad \text{for } l \in \mathcal{N}_k \setminus C(k), \tag{6}$$

where $\gamma \in [0,1]$ is a positive constant (note that in $\sigma_{kl}^2(n)$, $\mathbf{w}_l(n)$ is used with the data for k^{th} node, namely, $d_k(n)$ and $\mathbf{u}_k(n)$). Then, defining $\rho_{\delta_{kl}}(n) = \rho_{kl}\delta_{kl}(n)$, the clustered multitask diffusion APA (2) is modified as

Adaptation:

$$\boldsymbol{\psi}_k'(n+1) = \mathbf{w}_k(n) + \mu\mathbf{U}_k(n)\big(\epsilon\mathbf{I}_P + \mathbf{U}_k^T(n)\mathbf{U}_k(n)\big)^{-1}\mathbf{e}_k(n),$$

Combination (inter-cluster):

$$\boldsymbol{\psi}_k(n+1) = \boldsymbol{\psi}_k'(n+1) + \mu\eta \sum_{l\in\mathcal{N}_k\setminus C(k)} \rho_{\delta_{kl}(n)} (\mathbf{w}_l(n) - \mathbf{w}_k(n)),$$

Combination (intra-cluster):

$$\mathbf{w}_k(n+1) = \sum_{l\in\mathcal{N}_k\cap C(k)} a_{lk}\boldsymbol{\psi}_l(n+1). \tag{7}$$

The following may then be observed in the context of the above:

- Suppose, at index n during the transient stage, node l performs poorly whereas node k performs reasonably well. This means $\sigma_l^2(n)$ and thus $\sigma_{kl}^2(n)$ are larger than $\sigma_k^2(n)$. From (5), we then have $\delta_{kl}(n) = 0$, and thus, node k stops learning from node l. On the other hand, as $\mathbf{w}_l^* \sim \mathbf{w}_k^*$, $\sigma_{lk}^2(n)$ will be less than $\sigma_l^2(n)$, and thus, from (5), $\delta_{lk}(n) = 1$, meaning node l will continue to learn from node k.
- In the steady-state when both $\sigma_k^2(n)$ and $\sigma_l^2(n)$ are at par, $\sigma_{kl}^2(n) > \sigma_k^2(n)$, simultaneously with $\sigma_{lk}^2(n) > \sigma_l^2(n)$. This means both $\delta_{kl}(n) = 0$ and $\delta_{lk}(n) = 0$, and the two nodes, k and l, stop learning from each other, or, equivalently, they stop pulling each other towards their respective optimal values. This obviously improves the overall convergence performance.

Since the control variable $\delta_{kl}(n) \in \{0,1\}$, we always have $0 \le \rho_{\delta_{kl}}(n) \le \rho_{kl}$. Also note that $\delta_{kk}(n) = 1$ if $\mathcal{N}_k \setminus C(k) = \emptyset$. At each node k, the proposed multitask diffusion APA incurs a small amount of additional computational overhead, i.e., $|\mathcal{N}_k/\mathcal{C}(k)|(L+3)$ extra multiplications and $|\mathcal{N}_k/\mathcal{C}(k)|(L+3)$ extra additions, where $|\mathcal{N}_k/\mathcal{C}(k)|$ denotes the number of inter-cluster neighboring nodes. Furthermore, each node requires $(|\mathcal{N}_k/\mathcal{C}(k)|+1)$ memory locations to store the error variance defined in (6). This slight increase in overhead is quite acceptable in view of the improvement in performance achieved.

III. Performance Analysis

In this section, we focus on the convergence behavior of the proposed improved clustered multitask diffusion APA, particularly with regard to the control variable introduced, and obtain expressions for the instantaneous as well as steady-state MSD by following the lines of the energy conservation approach [23].

A. Network Global Model

Before proceeding to analysis, we define the network level optimal filter coefficient vector $\mathbf{w}^\star$, estimated filter coefficient vector $\mathbf{w}(n)$, input data matrix $\mathbf{U}(n)$ and the observation noise vector $\boldsymbol{\vartheta}(n)$ as follows:

$$\begin{aligned}\mathbf{w}^\star &= \text{col}\{\mathbf{w}_1^\star, \mathbf{w}_2^\star, \cdots, \mathbf{w}_N^\star\},\\ \mathbf{w}(n) &= \text{col}\{\mathbf{w}_1(n), \mathbf{w}_2(n), \cdots, \mathbf{w}_N(n)\},\\ \boldsymbol{\vartheta}(n) &= \text{col}\{\boldsymbol{\vartheta}_1(n), \boldsymbol{\vartheta}_2(n), \cdots, \boldsymbol{\vartheta}_N(n)\},\\ \mathbf{U}(n) &= \text{blockdiag}\{\mathbf{U}_1(n), \mathbf{U}_2(n), \cdots, \mathbf{U}_N(n)\},\end{aligned} \tag{8}$$

where col{·} and blockdiag{·} are used to denote the column wise stacking operator and block diagonalization operator, respectively. From the above definitions, the global data model is given by

$$\mathbf{d}(n) = \text{col}\{\mathbf{d}_1(n), \mathbf{d}_2(n), \cdots, \mathbf{d}_N(n)\} = \mathbf{U}^T(n)\mathbf{w}^\star + \boldsymbol{\vartheta}(n), \tag{9}$$

and the global error vector is given by

$$\begin{aligned}\mathbf{e}(n) &= \text{col}\{\mathbf{e}_1(n), \mathbf{e}_2(n), \cdots, \mathbf{e}_N(n)\}\\ &= \mathbf{d}(n) - \mathbf{U}^T(n)\mathbf{w}(n).\end{aligned} \tag{10}$$

Using the above definitions, from (7), the global model of the proposed improved clustered multitask diffusion APA can then be described as

$$\mathbf{w}(n+1) = \boldsymbol{\mathcal{A}}\big(\mathbf{w}(n) + \mu\mathbf{U}(n)(\epsilon\mathbf{I}_{PN} + \mathbf{U}^T(n)\mathbf{U}(n))^{-1}\mathbf{e}(n)\big), - \mu\eta\boldsymbol{\mathcal{A}}\boldsymbol{\mathcal{Q}}_\delta(n)\mathbf{w}(n), \tag{11}$$

where

$$\begin{aligned}\boldsymbol{\mathcal{A}} &= \mathbf{A}^T \otimes \mathbf{I}_L\\ \boldsymbol{\mathcal{Q}}_\delta(n) &= (\mathbf{D}_\delta(n) \otimes \mathbf{I}_L) - (\mathbf{P}_\delta(n) \otimes \mathbf{I}_L),\end{aligned} \tag{12}$$

with

$$\begin{aligned}\mathbf{D}_\delta(n) &= \text{diag}\{\rho_{\delta_1}(n), \rho_{\delta_2}(n), \cdots, \rho_{\delta_N}(n)\},\\ \mathbf{P}_{\boldsymbol{\delta}}(n) &= \mathbf{P} \odot \boldsymbol{\delta}(n),\end{aligned} \tag{13}$$

and

$$\rho_{\delta_k}(n) = \sum_{l\in\mathcal{N}_k\setminus C(k)} \rho_{\delta_{kl}}(n) = \sum_{l\in\mathcal{N}_k\setminus C(k)} \rho_{kl}\delta_{kl}(n) \quad \text{for } k = 1, \cdots, N. \tag{14}$$

The term $\boldsymbol{\delta}(n)$ denotes a $N\times N$ matrix with $[\boldsymbol{\delta}(n)]_{k,l} = \delta_{kl}(n)$, $\mathbf{A}$ is a $N\times N$ symmetric left stochastic matrix with $[\mathbf{A}]_{l,k} = a_{lk}$, and $\mathbf{P}$ is a $N\times N$ asymmetric right stochastic matrix with $[\mathbf{P}]_{k,l} = \rho_{kl}$. The symbols $\otimes$ and $\odot$ denote the right Kronecker product and Hadamard product operators, respectively [25]. In the following, we study the convergence behavior of the proposed improved clustered multitask diffusion APA given by (11).

In (7), every node is influenced by the local information (i.e., current and past data of the node itself which is temporal) as well as the information coming from neighbors through diffusion mode of cooperation (which is for the current cycle and is spatial). This simultaneous presence of spatial and temporal structures makes the analysis of distributed adaptive filters more challenging compared to single adaptive filters. In order to circumvent this difficulty, we assume the following:

Assumption 1: The data signal $u_k(n)$ arise from a stationary random process that is temporally stationary with the

correlation matrix $\mathbf{R}_{u,k} = E[\mathbf{u}_k(n)\mathbf{u}_k^T(n)]$ and the data matrices $\mathbf{U}_k(n)$, $k = 1, 2, \cdots N$ are spatially independent.

Assumption 2: Observation noise $\vartheta_k(n)$ is taken to be spatially and temporally i.i.d. Gaussian with mean zero and variance $\sigma^2_{\vartheta,k}$.

Assumption 3: The network topology is assumed to be static, meaning the combiner coefficients are constant throughout the process.

Assumption 4: We assume statistical independence between $\mathbf{w}_k(n)$ and $\mathbf{U}_k(n)$, $\mathbf{d}_k(n)$, $k = 1, 2, \cdots, N$ (i.e., generalized independence assumption), implying that $\mathbf{w}_k(n)$ is also statistical independent of $\boldsymbol{\vartheta}_k(n)$. Further, $\delta_{kl}(n)$ is considered to be statistically independent of $\mathbf{U}_l(n)$, $\vartheta_l(n)$ and $\mathbf{w}_l(n)$, $k, l = 1, 2, \cdots, N$.

Assumption 5: The step size μ is sufficiently small so that the terms involving higher order powers of μ can be ignored.

The above assumptions are commonly used in the analysis of diffusion adaptive strategies. Apart from these, the analysis also requires properties of the block maximum norm of a matrix (i.e., $\|\cdot\|_{b,\infty}$), the block vectorization operator (i.e., bvec$\{\cdot\}$), and the block Kronecker product of two matrices (i.e., $\otimes_b$). For convenience of the reader, we briefly summarize below the definition and properties of the above. Details of the same can be found in [3].

Firstly, given a block column vector $\mathbf{x} = \text{col}\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$, where the individual entries $\mathbf{x}_k$, $k = 1, 2, \cdots, N$ are $L \times 1$ vectors, the block maximum norm of $\mathbf{x}$ is defined as

$$\|\mathbf{x}\|_{b,\infty} \triangleq \max_{1\le k\le N} \|\mathbf{x}_k\|_2, \tag{15}$$

where $\|\cdot\|_2$ denotes the Euclidean norm of its vector argument. Correspondingly, the block maximum norm of an arbitrary block matrix $\boldsymbol{\mathcal{A}}$ whose individual block entries of size $L \times L$ each, is defined as

$$\|\boldsymbol{\mathcal{A}}\|_{b,\infty} \triangleq \max_{\mathbf{x}\neq 0} \frac{\|\boldsymbol{\mathcal{A}}\mathbf{x}\|_{b,\infty}}{\|\mathbf{x}\|_{b,\infty}}. \tag{16}$$

The block maximum norm of a block matrix satisfies all the standard properties of a matrix norm, namely, (i) non-negativity (i.e., $\|\boldsymbol{\mathcal{A}}\|_{b,\infty}$ is real and non-negative, and equals zero iff $\boldsymbol{\mathcal{A}}$ is an all-zero matrix), (ii) homogeneity (i.e., for any scalar c, $\|c\boldsymbol{\mathcal{A}}\|_{b,\infty} = |c|\|\boldsymbol{\mathcal{A}}\|_{b,\infty}$), (iii) triangle inequality (i.e., $\|\boldsymbol{\mathcal{A}}+\boldsymbol{\mathcal{B}}\|_{b,\infty} \le \|\boldsymbol{\mathcal{A}}\|_{b,\infty} + \|\boldsymbol{\mathcal{B}}\|_{b,\infty}$), and sub-multiplicativity (i.e., $\|\boldsymbol{\mathcal{A}}\boldsymbol{\mathcal{B}}\|_{b,\infty} \le \|\boldsymbol{\mathcal{A}}\|_{b,\infty}\|\boldsymbol{\mathcal{B}}\|_{b,\infty}$). Further, given a block matrix $\boldsymbol{\mathcal{A}}$, its spectral radius $\rho(\boldsymbol{\mathcal{A}})$ (i.e., magnitude of the largest (in magnitude) eigenvalue) is bounded by its block maximum norm, i.e., $\rho(\boldsymbol{\mathcal{A}}) \le \|\boldsymbol{\mathcal{A}}\|_{b,\infty}$.

In addition to the above, the block maximum norm has a few other useful properties given as follows [1]:

1) Let $\mathbf{A}$ be an $N \times N$ matrix with bounded entries and let $\boldsymbol{\mathcal{A}} = \mathbf{A} \otimes \mathbf{I}_L$. Then $\|\boldsymbol{\mathcal{A}}\|_{b,\infty} = \|\mathbf{A}\|_\infty$, where $\|\cdot\|_\infty$ denotes the maximum absolute row sum of the argument matrix, i.e., $\|\mathbf{A}\|_\infty = \max_{1\le i\le N} \sum_{j=1}^{N} |a_{ij}|$ (where $[\mathbf{A}]_{i,j} = a_{ij}$).
2) Let $\boldsymbol{\mathcal{D}} = \text{diag}\{\mathbf{D}_1, \mathbf{D}_2, \cdots, \mathbf{D}_N\}$ be a block diagonal matrix, where each $\mathbf{D}_k$, for $k = 1, 2, \cdots, N$, is a Hermitian matrix of size $L \times L$. Then it holds that $\rho(\boldsymbol{\mathcal{D}}) = \max_{1\le k\le N} \rho(\mathbf{D}_k) = \|\boldsymbol{\mathcal{D}}\|_{b,\infty}$.

Next, let $\boldsymbol{\Sigma}$ be a $LN \times LN$ block matrix, with $(i, j)^{th}$ block (of size $L \times L$) given by $\boldsymbol{\Sigma}_{ij}$ $i, j = 1, 2, \cdots, N$. Then, bvec$\{\boldsymbol{\Sigma}\}$ produces an $L^2 N^2 \times 1$ vector $\boldsymbol{\sigma}$ from $\boldsymbol{\Sigma}$ as $\boldsymbol{\sigma} = \text{bvec}\{\boldsymbol{\Sigma}\} = \text{vec}\{\boldsymbol{\sigma}_1, \boldsymbol{\sigma}_2, \cdots, \boldsymbol{\sigma}_N\}$, with $\boldsymbol{\sigma}_j = \text{vec}\{\text{vec}\{\boldsymbol{\Sigma}_{1j}\}, \text{vec}\{\boldsymbol{\Sigma}_{2j}\}, \cdots, \text{vec}\{\boldsymbol{\Sigma}_{Nj}\}\}$, $j = 1, 2, \cdots, N$ ("vec$\{\cdot\}$" is the so-called vectorization operator that stacks successive columns of the argument matrix downwards, starting from the leftmost column). Clearly, bvec$\{\cdot\}$ is a one-to-one operator, meaning, given $\boldsymbol{\sigma} = \text{bvec}\{\boldsymbol{\Sigma}\}$, we can also define $\boldsymbol{\Sigma} = \text{bvec}^{-1}\{\boldsymbol{\sigma}\}$.

Lastly, the block Kronecker product of any two block matrices $\mathbf{A}$ and $\mathbf{B}$ of size $LN \times LN$ with $L \times L$ block element matrices $\mathbf{A}_{ij}$ for $i, j = 1, 2, \cdots, N$ and $\mathbf{B}_{kl}$ for $k, l = 1, 2, \cdots, N$ is denoted by $\mathbf{A} \otimes_b \mathbf{B}$, and is given by a $L^2N^2 \times L^2N^2$ block matrix, with $(i, j)^{th}$ block given by $\mathbf{A}_{ij} \otimes_b \mathbf{B}$, $i, j = 1, 2, \cdots, N$, where, $\mathbf{A}_{ij} \otimes_b \mathbf{B}$ is again a block matrix of size $L^2N \times L^2N$, with the $(k, l)^{th}$ entry given by $\mathbf{A}_{ij} \otimes \mathbf{B}_{kl}$, $k, l = 1, 2, \cdots, N$. Given the following block matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$ of size $LN \times LN$ each with $L \times L$ block element matrices, the following properties hold [3]:

$$(\mathbf{A}+\mathbf{B}) \otimes_b (\mathbf{C}+\mathbf{D}) = (\mathbf{A} \otimes_b \mathbf{C}) + (\mathbf{A} \otimes_b \mathbf{D}) + (\mathbf{B} \otimes_b \mathbf{C}) + (\mathbf{B} \otimes_b \mathbf{D}), \tag{17a}$$

$$(\mathbf{AC} \otimes_b \mathbf{BD}) = (\mathbf{A} \otimes_b \mathbf{B})(\mathbf{C} \otimes_b \mathbf{D}), \tag{17b}$$

$$(\mathbf{A} \otimes \mathbf{B}) \otimes_b (\mathbf{C} \otimes \mathbf{D}) = (\mathbf{A} \otimes \mathbf{C}) \otimes (\mathbf{B} \otimes \mathbf{D}), \tag{17c}$$

$$Tr(\mathbf{AB}) = (\text{bvec}\{\mathbf{A}^T\})^T \text{bvec}\{\mathbf{B}\}, \tag{17d}$$

$$(\mathbf{A} \otimes_b \mathbf{B})^T = (\mathbf{A}^T \otimes_b \mathbf{B}^T), \tag{17e}$$

$$\text{bvec}\{\mathbf{B\Sigma A}^T\} = (\mathbf{A} \otimes_b \mathbf{B})\text{bvec}\{\boldsymbol{\Sigma}\}, \tag{17f}$$

$$\lambda(\mathbf{A} \otimes_b \mathbf{B}) = \{\lambda_i(\mathbf{A})\lambda_j(\mathbf{B})\}_{i=1,j=1}^{LN,LN}, \tag{17g}$$

where $Tr(\cdot)$ denotes the trace of its argument matrix and $\lambda_i(\cdot)$ is the i^{th} eigenvalue of its argument matrix.

B. First Order Convergence Analysis

Denoting the global weight deviation vector of the improved clustered multitask diffusion APA at n-th index as $\widetilde{\mathbf{w}}(n) = \mathbf{w}^\star - \mathbf{w}(n)$, recalling $\boldsymbol{\mathcal{A}}\mathbf{w}^\star = \mathbf{w}^\star$, from (11), the recursion for $\widetilde{\mathbf{w}}(n)$ can then be obtained as

$$\widetilde{\mathbf{w}}(n+1) = \boldsymbol{\mathcal{B}}_\delta(n)\widetilde{\mathbf{w}}(n) - \mu\boldsymbol{\mathcal{A}}\mathbf{U}(\boldsymbol{n}) (\epsilon\mathbf{I}_{PN} + \mathbf{U}^T(n)\mathbf{U}(n))^{-1}\boldsymbol{\vartheta}(n) + \mathbf{r}_\delta(n), \tag{18}$$

where $\boldsymbol{\mathcal{B}}_\delta(n) = \boldsymbol{\mathcal{A}}(\mathbf{I}_{LN} - \mu\mathbf{Z}(n) - \mu\eta\boldsymbol{\mathcal{Q}}_\delta(n))$, $\mathbf{Z}(n) = \mathbf{U}(n)(\epsilon\mathbf{I}_{PN} + \mathbf{U}^T(n)\mathbf{U}(n))^{-1}\mathbf{U}^T(n)$ and $\mathbf{r}_\delta(n) = \mu\eta\boldsymbol{\mathcal{A}}\boldsymbol{\mathcal{Q}}_\delta(n)\mathbf{w}^\star$. In the following, we establish the condition of convergence of the improved clustered multitask diffusion APA, i.e., (7), where we use the definition $\overline{\mathbf{Z}}_k = E[\mathbf{Z}_k(n)] = E[\mathbf{U}_k(n)(\epsilon\mathbf{I}_P + \mathbf{U}_k^T(n)\mathbf{U}_k(n))^{-1}\mathbf{U}_k^T(n)]$ (the index n is dropped from $\overline{\mathbf{Z}}_k$ due to stationarity of $u_k(n)$).

Theorem 1: Assume the data model (9) and the assumptions 1-4 to hold (assumption 5 not needed here). Then a sufficient condition for the improved clustered multitask diffusion APA to converge in mean is

$$0 < \mu < \frac{2}{\max\limits_{1\le k\le N}\{\max\limits_{1\le i\le L}\{\lambda_i(\overline{\mathbf{Z}}_k)\}\} + 2\eta \max\limits_{1\le k\le N}\{E[\rho_{\delta_k}(n)]\}}, \tag{19}$$

where $E[\rho_{\delta_k}(n)] = \sum_{l\in\mathcal{N}_k\setminus C(k)} E[\rho_{\delta_{kl}}(n)] = \sum_{l\in\mathcal{N}_k\setminus C(k)} \rho_{kl} E[\delta_{kl}(n)]$.

Proof: Given in the Appendix A. □

Recalling that $\delta_{kl}(n) \in \{0, 1\}$, implying $0 \leq E[\delta_{kl}(n)] \leq 1$, we have $E[\rho_{\delta_k}(n)] = \sum_{l \in \mathcal{N}_k \backslash C(k)} \rho_{kl} E[\delta_{kl}(n)] \leq 1$, $k = 1, 2, \cdots N$. Clearly, the presence of the control variable $\delta_{kl}(n)$ in the improved clustered multitask diffusion strategy results in a larger upper bound of μ for convergence in mean as compared to the conventional clustered multitask diffusion APA [22]. Under (19), letting $n \to \infty$ on both sides of (39), we then have

$$\lim_{n\to\infty} E[\widetilde{\mathbf{w}}(n)] = (\mathbf{I}_{LN} - E[\boldsymbol{\mathcal{B}}_\delta(\infty)])^{-1} E[\mathbf{r}_\delta(\infty)], \quad (20)$$

where $E[\boldsymbol{\mathcal{B}}_\delta(\infty)] = \lim_{n\to\infty} E[\boldsymbol{\mathcal{B}}_\delta(n)]$ and $E[\mathbf{r}_\delta(\infty)] = \lim_{n\to\infty} E[\mathbf{r}_\delta(n)]$.

C. Second Order Convergence Analysis

Defining the weighted norm-square of $\widetilde{\mathbf{w}}(n)$ as $\|\widetilde{\mathbf{w}}(n)\|^2_{\boldsymbol{\Sigma}} = \widetilde{\mathbf{w}}(n)^T \boldsymbol{\Sigma} \widetilde{\mathbf{w}}(n)$ where $\boldsymbol{\Sigma}$ is a positive semi-definite matrix that can be chosen arbitrarily, and using assumption $1-4$, one can write from (18),

$$\begin{aligned} &E\big[\|\widetilde{\mathbf{w}}(n+1)\|^2_{\boldsymbol{\Sigma}}\big] \\ &\quad = E\big[\|\widetilde{\mathbf{w}}(n)\|^2_{\boldsymbol{\Sigma}_\delta(n)}\big] + \mu^2 E\big[\boldsymbol{\vartheta}^T(n)\mathbf{Y}^{\boldsymbol{\Sigma}}(n)\boldsymbol{\vartheta}(n)\big] \\ &\qquad + E\big[\|\mathbf{r}_\delta(n)\|^2_{\boldsymbol{\Sigma}}\big] + E\big[\widetilde{\mathbf{w}}^T(n)\boldsymbol{\mathcal{B}}^T_\delta(n)\boldsymbol{\Sigma}\mathbf{r}_\delta(n)\big] \\ &\qquad + E\big[\mathbf{r}^T_\delta(n)\boldsymbol{\Sigma}\boldsymbol{\mathcal{B}}_\delta(n)\widetilde{\mathbf{w}}(n)\big], \end{aligned} \quad (21)$$

where the cross terms turn out to be zero as $\boldsymbol{\vartheta}_k(n)$ is taken to be zero mean and statistically independent of $\mathbf{U}_l(m)$, for all k, l and m, n, and as $\mathbf{w}_k(n)$ and $\delta_{kl}(n)$ are statistically independent of $\boldsymbol{\vartheta}_k(n)$, $k, l = 1, 2, \cdots, N$. The weighted matrix $\boldsymbol{\Sigma}_\delta(n)$ is given by

$$\boldsymbol{\Sigma}_\delta(n) = E[\boldsymbol{\mathcal{B}}^T_\delta(n)\boldsymbol{\Sigma}\boldsymbol{\mathcal{B}}_\delta(n)], \quad (22)$$

and

$$\mathbf{Y}^{\boldsymbol{\Sigma}}(n) = \mathbf{X}(n)\boldsymbol{\mathcal{A}}^T\boldsymbol{\Sigma}\boldsymbol{\mathcal{A}}\mathbf{X}^T(n), \quad (23)$$

with $\mathbf{X}(n) = \big(\epsilon\mathbf{I}_{PN} + \mathbf{U}^T(n)\mathbf{U}(n)\big)^{-1}\mathbf{U}^T(n)$. Using (17f), the vectors $\boldsymbol{\sigma} = \text{bvec}\{\boldsymbol{\Sigma}\}$ and $\boldsymbol{\sigma}_\delta(n) = \text{bvec}\{\boldsymbol{\Sigma}_\delta(n)\}$ can be related as

$$\boldsymbol{\sigma}_\delta(n) = \boldsymbol{\mathcal{F}}^T_\delta(n)\boldsymbol{\sigma}, \quad (24)$$

where

$$\boldsymbol{\mathcal{F}}_\delta(n) = E[\boldsymbol{\mathcal{B}}_\delta(n) \otimes_b \boldsymbol{\mathcal{B}}_\delta(n)] = \big(\boldsymbol{\mathcal{A}} \otimes_b \boldsymbol{\mathcal{A}}\big)\boldsymbol{\mathcal{H}}_\delta(n), \quad (25)$$

with

$$\begin{aligned} \boldsymbol{\mathcal{H}}_\delta(n) \approx\ & \mathbf{I}_{L^2N^2} - \mu(\mathbf{I}_{LN} \otimes_b \overline{\mathbf{Z}}) - \mu(\overline{\mathbf{Z}} \otimes_b \mathbf{I}_{LN}) \\ & -\mu\eta(E[\boldsymbol{\mathcal{Q}}_\delta(n)] \otimes_b \mathbf{I}_{LN}) - \mu\eta(\mathbf{I}_{LN} \otimes_b E[\boldsymbol{\mathcal{Q}}_\delta(n)]). \end{aligned} \quad (26)$$

In the above, under the assumption 5, the terms involving higher order moments of μ are ignored and we continue our analysis with this approximation.

Next, we consider the second term on the R.H.S of (21). we can write $E[\boldsymbol{\vartheta}^T(n)\mathbf{Y}^{\boldsymbol{\Sigma}}(n)\boldsymbol{\vartheta}(n)\big] = E[\boldsymbol{\vartheta}^T(n)\mathbf{X}(n)\boldsymbol{\mathcal{A}}^T\boldsymbol{\Sigma}\boldsymbol{\mathcal{A}}\mathbf{X}^T(n)\boldsymbol{\vartheta}(n)] = E[Tr(\boldsymbol{\vartheta}^T(n)\mathbf{X}(n)\boldsymbol{\mathcal{A}}^T\boldsymbol{\Sigma}\boldsymbol{\mathcal{A}}\mathbf{X}^T(n)\boldsymbol{\vartheta}(n))] = Tr(\boldsymbol{\mathcal{A}}E[\mathbf{X}^T(n)\boldsymbol{\vartheta}(n)\boldsymbol{\vartheta}^T(n)\mathbf{X}(n)]\boldsymbol{\mathcal{A}}^T\boldsymbol{\Sigma})$. Using the spatial and temporal whiteness of $\vartheta_k(n)$, and recalling $\boldsymbol{\vartheta}_k(n)$ is statistically independent of $\mathbf{U}_l(m)$, for all k, l and m, n, we can then obtain

$$\begin{aligned} &Tr\big(\boldsymbol{\mathcal{A}}E[\mathbf{X}^T(n)\boldsymbol{\vartheta}(n)\boldsymbol{\vartheta}^T(n)\mathbf{X}(n)]\boldsymbol{\mathcal{A}}^T\boldsymbol{\Sigma}\big) \\ &\qquad = Tr(\boldsymbol{\mathcal{A}}E[\boldsymbol{\Phi}(n)]\boldsymbol{\mathcal{A}}^T\boldsymbol{\Sigma}), \end{aligned} \quad (27)$$

where $\boldsymbol{\Phi}(n) = \mathbf{X}^T(n)\boldsymbol{\Lambda}_\vartheta\mathbf{X}(n)$, with $\boldsymbol{\Lambda}_\vartheta = E[\boldsymbol{\vartheta}(n)\boldsymbol{\vartheta}^T(n)] = \text{diag}\{\sigma^2_{\vartheta,1}\mathbf{I}_P, \sigma^2_{\vartheta,2}\mathbf{I}_P, \cdots, \sigma^2_{\vartheta,N}\mathbf{I}_P\}$, a $PN \times PN$ diagonal matrix. Using (17d), we finally have

$$Tr(\boldsymbol{\mathcal{A}}E[\boldsymbol{\Phi}(n)]\boldsymbol{\mathcal{A}}^T\boldsymbol{\Sigma}) = \boldsymbol{\gamma}^T\boldsymbol{\sigma}, \quad (28)$$

where

$$\boldsymbol{\gamma} = \text{bvec}\{\boldsymbol{\mathcal{A}}E[\boldsymbol{\Phi}(n)]\boldsymbol{\mathcal{A}}^T\} = (\boldsymbol{\mathcal{A}} \otimes \boldsymbol{\mathcal{A}})\boldsymbol{\gamma}_\vartheta, \quad (29)$$

with $\boldsymbol{\gamma}_\vartheta = \text{bvec}\{E[\mathbf{X}^T(n)\boldsymbol{\Lambda}_\vartheta\mathbf{X}(n)]\} = \text{bvec}\{\text{diag}(\sigma^2_{\vartheta,1} E[\mathbf{U}_1(n)(\epsilon\mathbf{I}_P + \mathbf{U}^T_1(n)\mathbf{U}_1(n))^{-2}\mathbf{U}^T_1(n)], \cdots, \sigma^2_{\vartheta,N}E[\mathbf{U}_N(n)(\epsilon\mathbf{I}_P + \mathbf{U}^T_N(n)\mathbf{U}_N(n))^{-2}\mathbf{U}^T_N(n)])\}$.

Lastly, we evaluate the last three terms on the R.H.S of (21). Firstly,

$$\begin{aligned} E\big[\|\mathbf{r}_\delta(n)\|^2_{\boldsymbol{\Sigma}}\big] = \mu^2\eta^2 Tr\big(&\boldsymbol{\mathcal{A}}E\big[\boldsymbol{\mathcal{Q}}_\delta(n)\mathbf{w}^\star(\mathbf{w}^\star)^T \\ &\boldsymbol{\mathcal{Q}}^T_\delta(n)\big]\boldsymbol{\mathcal{A}}^T\boldsymbol{\Sigma}\big). \end{aligned} \quad (30)$$

From (17d), we get

$$Tr(\boldsymbol{\mathcal{A}}E[\boldsymbol{\mathcal{Q}}_\delta(n)\mathbf{w}^\star(\mathbf{w}^\star)^T\boldsymbol{\mathcal{Q}}^T_\delta(n)]\boldsymbol{\mathcal{A}}^T\boldsymbol{\Sigma}) = \mu^2\eta^2\mathbf{r}^T_{b,\delta}(n)\boldsymbol{\sigma}, \quad (31)$$

where

$$\begin{aligned} \mathbf{r}_{b,\delta}(n) &= \text{bvec}\{\boldsymbol{\mathcal{A}}E[\boldsymbol{\mathcal{Q}}_\delta(n)\mathbf{w}^\star(\mathbf{w}^\star)^T\boldsymbol{\mathcal{Q}}^T_\delta(n)]\boldsymbol{\mathcal{A}}^T\} \\ &= (\boldsymbol{\mathcal{A}} \otimes_b \boldsymbol{\mathcal{A}})E[\boldsymbol{\mathcal{Q}}_\delta(n) \otimes_b \boldsymbol{\mathcal{Q}}_\delta(n)]\text{bvec}\{\mathbf{w}^\star(\mathbf{w}^\star)^T\}. \end{aligned} \quad (32)$$

Next, we consider the term $E[\widetilde{\mathbf{w}}^T(n)\boldsymbol{\mathcal{B}}^T_\delta(n)\boldsymbol{\Sigma}\mathbf{r}_\delta(n)]$ on the R.H.S of (21). Recalling that the block vectorization of a scalar results in itself, we can write

$$\begin{aligned} &E[\widetilde{\mathbf{w}}^T(n)\boldsymbol{\mathcal{B}}^T_\delta(n)\boldsymbol{\Sigma}\mathbf{r}_\delta(n)] \\ &\quad = E[\text{bvec}\big\{\widetilde{\mathbf{w}}^T(n)\boldsymbol{\mathcal{B}}^T_\delta(n)\boldsymbol{\Sigma}\mathbf{r}_\delta(n)\big\}\big] \\ &\quad = E\big[\mathbf{r}_\delta(n) \otimes_b \boldsymbol{\mathcal{B}}_\delta(n)\widetilde{\mathbf{w}}(n)\big]^T\boldsymbol{\sigma} \\ &\quad = \big(E[\mathbf{r}_\delta(n) \otimes_b \boldsymbol{\mathcal{B}}_\delta(n)](1 \otimes_b E[\widetilde{\mathbf{w}}(n)])\big)^T\boldsymbol{\sigma} \\ &\quad = E[\widetilde{\mathbf{w}}^T(n)]\boldsymbol{\alpha}^T_\delta(n)\boldsymbol{\sigma}, \end{aligned} \quad (33)$$

where $\boldsymbol{\alpha}_\delta(n) = E[\mathbf{r}_\delta(n) \otimes_b \boldsymbol{\mathcal{B}}_\delta(n)] = (\boldsymbol{\mathcal{A}} \otimes_b \boldsymbol{\mathcal{A}})\big(\mu\eta(E[\boldsymbol{\mathcal{Q}}_\delta(n)]\mathbf{w}^\star \otimes_b \mathbf{I}_{LN}) - \mu^2\eta(E[\boldsymbol{\mathcal{Q}}_\delta(n)]\mathbf{w}^\star \otimes_b \overline{\mathbf{Z}}) - \mu^2\eta^2 E[\boldsymbol{\mathcal{Q}}_\delta(n) \otimes_b \boldsymbol{\mathcal{Q}}_\delta(n)](\mathbf{w}^\star \otimes_b \mathbf{I}_{LN})\big) \approx \mu\eta(\boldsymbol{\mathcal{A}} \otimes_b \boldsymbol{\mathcal{A}})(E[\boldsymbol{\mathcal{Q}}_\delta(n)]\mathbf{w}^\star \otimes_b \mathbf{I}_{LN})$ (i.e., after neglecting terms having higher order powers of μ).

Finally, the last term on the R.H.S of (21), viz., $E[\mathbf{r}^T_\delta(n)\boldsymbol{\Sigma}\boldsymbol{\mathcal{B}}_\delta(n)\widetilde{\mathbf{w}}(n)]$ is easily seen to be the same as the previous term, $E[\widetilde{\mathbf{w}}^T(n)\boldsymbol{\mathcal{B}}^T_\delta(n)\boldsymbol{\Sigma}\mathbf{r}_\delta(n)]$ evaluated in (33), as both are scalars and one is the transpose of the other.

Combining all these together, the mean square of the weight deviation vector for the improved clustered multitask diffusion APA can then be obtained as

$$\begin{aligned} &E[\|\widetilde{\mathbf{w}}(n+1)\|^2_{\text{bvec}^{-1}\{\boldsymbol{\sigma}\}}] \\ &\quad = E[\|\widetilde{\mathbf{w}}(n)\|^2_{\text{bvec}^{-1}\{\boldsymbol{\mathcal{F}}^T_\delta(n)\boldsymbol{\sigma}\}}] \\ &\qquad + \mu^2\boldsymbol{\gamma}^T\boldsymbol{\sigma} + \mathbf{f}\big(\mathbf{r}_{b,\delta}(n), \boldsymbol{\alpha}_\delta(n), E[\widetilde{\mathbf{w}}(n)], \boldsymbol{\sigma}\big), \end{aligned} \quad (34)$$

where $\mathbf{f}\big(\mathbf{r}_{b,\delta}(n), \boldsymbol{\alpha}_\delta(n), E[\widetilde{\mathbf{w}}(n)], \boldsymbol{\sigma}\big) = \mu^2\eta^2\mathbf{r}^T_{b,\delta}(n)\boldsymbol{\sigma} + 2E[\widetilde{\mathbf{w}}^T(n)]\boldsymbol{\alpha}^T_\delta(n)\boldsymbol{\sigma}$.

Theorem 2: Assume the data model (9) and the assumptions 1-5 to hold. Assume further that the step size is sufficiently small such that the approximation (26) is justified by ignoring the higher order powers of the step size, and (34) can be used as a reasonable representation for studying the dynamics of the weighted MSD. Then, the improved clustered multitask diffusion APA exhibits stable MSD performance under

$$0<\mu<\frac{1}{\max\limits_{1\le k\le N}\left\{\max\limits_{1\le i\le L}\{\lambda_i(\overline{\mathbf{Z}}_k)\}\right\}+2\eta\max\limits_{1\le k\le N}\left\{E[\rho_{\delta_k}(n)]\right\}}, \tag{35}$$

Proof: Given in the Appendix B. □

Under (35), letting $n\to\infty$ on both sides of (34), we will have

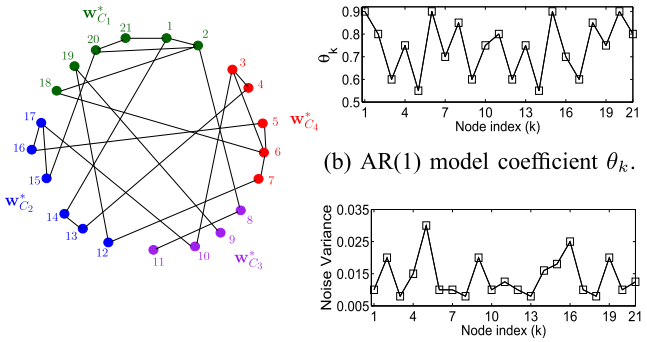
$$\lim_{n\to\infty}E[\|\widetilde{\mathbf{w}}(n)\|^2_{\text{bvec}^{-1}\{(\mathbf{I}_{L^2N^2}-\boldsymbol{\mathcal{F}}_\delta^T(n))\boldsymbol{\sigma}\}}] = \mu^2\boldsymbol{\gamma}^T\boldsymbol{\sigma}+\mathbf{f}\left(\mathbf{r}_{b,\delta}(\infty),\boldsymbol{\alpha}_\delta(\infty),E[\widetilde{\mathbf{w}}(\infty)],\boldsymbol{\sigma}\right), \tag{36}$$

where $\mathbf{r}_{b,\delta}(\infty)=\lim\limits_{n\to\infty}\mathbf{r}_{b,\delta}(n)$, $\boldsymbol{\alpha}_\delta(\infty)=\lim\limits_{n\to\infty}\boldsymbol{\alpha}_\delta(n)$ and $E[\widetilde{\mathbf{w}}(\infty)]=\lim\limits_{n\to\infty}E[\widetilde{\mathbf{w}}(n)]$. Since $\|\boldsymbol{\mathcal{F}}_\delta^T(\infty)\|_{b,\infty}<1$ (under the convergence condition (35), as shown in Appendix B), where $\boldsymbol{\mathcal{F}}_\delta(\infty)=\lim\limits_{n\to\infty}\boldsymbol{\mathcal{F}}_\delta(n)$, the matrix $(\mathbf{I}_{L^2N^2}-\boldsymbol{\mathcal{F}}_\delta^T(\infty))$ is invertible. By choosing $\boldsymbol{\sigma}=\frac{1}{N}\left(\mathbf{I}_{L^2N^2}-\boldsymbol{\mathcal{F}}_\delta^T(\infty)\right)^{-1}\text{bvec}\{\mathbf{I}_{LN}\}$, network level, steady-state MSD of the improved clustered multitask diffusion APA, i.e., $\xi(\infty)=\frac{1}{N}\lim\limits_{n\to\infty}E[\|\widetilde{\mathbf{w}}(n)\|^2]$ can be obtained as follows:

$$\xi(\infty)=\frac{1}{N}\mu^2\boldsymbol{\gamma}^T\left(\mathbf{I}_{L^2N^2}-\boldsymbol{\mathcal{F}}_\delta^T(\infty)\right)^{-1}\text{bvec}\{\mathbf{I}_{LN}\}+\mathbf{f}\left(\begin{matrix}\mathbf{r}_{b,\delta}(\infty),\boldsymbol{\alpha}_\delta(\infty),E[\widetilde{\mathbf{w}}(\infty)],\\ \frac{1}{N}(\mathbf{I}_{L^2N^2}-\boldsymbol{\mathcal{F}}_\delta^T(\infty))^{-1}\text{bvec}\{\mathbf{I}_{LN}\}\end{matrix}\right). \tag{37}$$

To relate network level, instantaneous MSD $\xi(n+1)$ recursively with $\xi(n)$, where $\xi(n)=\frac{1}{N}E[\|\widetilde{\mathbf{w}}(n)\|^2]$, using (46), $E[\|\widetilde{\mathbf{w}}(n+1)\|^2_{\boldsymbol{\Sigma}}]=E[\|\widetilde{\mathbf{w}}(n+1)\|^2_{\text{bvec}^{-1}\{\boldsymbol{\sigma}\}}]$ and $[E\|\widetilde{\mathbf{w}}(n)\|^2_{\boldsymbol{\Sigma}}]=E[\|\widetilde{\mathbf{w}}(n)\|^2_{\text{bvec}^{-1}\{\boldsymbol{\sigma}\}}]$ can be related as given below:

$$\begin{aligned}
&E[\|\widetilde{\mathbf{w}}(n+1)\|^2_{\text{bvec}^{-1}\{\boldsymbol{\sigma}\}}]\\
&=E[\|\widetilde{\mathbf{w}}(n)\|^2_{\text{bvec}^{-1}\{\boldsymbol{\sigma}\}}]+\mu^2\boldsymbol{\gamma}^T\boldsymbol{\mathcal{F}}_\delta^T(n)\,\boldsymbol{\sigma}\\
&\quad+\mu^2\boldsymbol{\gamma}^T\left(\sum_{i=1}^{n-1}\left(\prod_{j=i}^{n-1}\boldsymbol{\mathcal{F}}_\delta^T(j)\right)\right)\left(\boldsymbol{\mathcal{F}}_\delta^T(n)\;-\mathbf{I}_{L^2N^2}\right)\boldsymbol{\sigma}\\
&\quad-E[\|\widetilde{\mathbf{w}}(0)\|^2_{\text{bvec}^{-1}\{\left(\prod_{i=0}^{n-1}\boldsymbol{\mathcal{F}}_\delta^T(i)\right)\left(\mathbf{I}_{L^2N^2}-\boldsymbol{\mathcal{F}}_\delta^T(n)\;\right)\boldsymbol{\sigma}\}}]\\
&\quad+\mathbf{f}(\mathbf{r}_{b,\delta}(n),\boldsymbol{\alpha}_\delta(n),E[\widetilde{\mathbf{w}}(n)],\boldsymbol{\sigma})\\
&\quad+\mathbf{f}\big(\mathbf{r}_{b,\delta}(n-1),\boldsymbol{\alpha}_\delta(n-1),E[\widetilde{\mathbf{w}}(n-1)],(\boldsymbol{\mathcal{F}}_\delta^T(n)-\mathbf{I}_{L^2N^2})\boldsymbol{\sigma}\big)\\
&\quad+\sum_{i=1}^{n-1}\mathbf{f}\left(\begin{matrix}\mathbf{r}_{b,\delta}(n-1-i),\boldsymbol{\alpha}_\delta(n-1-i),E[\widetilde{\mathbf{w}}(n-1-i)],\\ \left(\prod\limits_{j=1}^{i}\boldsymbol{\mathcal{F}}_\delta^T(n-j)\right)(\boldsymbol{\mathcal{F}}_\delta^T(n)-\mathbf{I}_{L^2N^2})\boldsymbol{\sigma}\end{matrix}\right).
\end{aligned} \tag{38}$$



(a) Network topology. (b) AR(1) model coefficient θ_k. (c) Noise variance $\sigma^2_{\vartheta_k}$.

Fig. 1. Network topology and node profile statistics.

By choosing $\boldsymbol{\Sigma}=\frac{1}{N}\mathbf{I}_{LN}$, implying $\boldsymbol{\sigma}=\frac{1}{N}\text{bvec}\{\mathbf{I}_{LN}\}$, the network level transient MSD, i.e., $\xi(n)=\frac{1}{N}E[\|\widetilde{\mathbf{w}}(n)\|^2]$ can be obtained.

IV. Simulation Studies

In this section, we demonstrate via numerical simulations the performance of the proposed improved clustered multitask diffusion APA. For this, we considered a clustered multitask network consisting of $N=21$ nodes with the topology shown in Fig. 1(a).

The nodes in the network were grouped into 4 clusters: $C_1=\{1,2,18,19,20,21\}$, $C_2=\{12,13,14,15,16,17\}$, $C_3=\{8,9,10,11\}$ and $C_4=\{3,4,5,6,7\}$. These clusters aim to estimate their respective 256 tap optimal parameter vectors in collaborative fashion which are chosen as $\mathbf{w}^*_{C_q}=\mathbf{w}_0+\delta_{C_q}\mathbf{w}_0$ for $q=1,2,3,4$ with $\delta_{C_1}=0,\delta_{C_2}=0.025$, $\delta_{C_3}=0.05$ and $\delta_{C_4}=0.075$. The coefficient vector $\mathbf{w}_0$ was generated from zero mean, unity variance Gaussian distribution. Simulations were conducted for colored Gaussian input of unit variance, where a unity variance colored, Gaussian input $u_k(n)$ was generated by driving the first order auto-regressive (AR) model: $u_k(n)=\theta_k u_k(n-1)+\sqrt{1-\theta_k^2}z_k(n)$, $|\theta_k|<1$ with a unity variance, white Gaussian input $z_k(n)$. The coefficient θ_k varies from node to node and its distribution against the node index k is shown in Fig. 1(b). The observation noise $\vartheta_k(n)$ was taken to be zero mean i.i.d. Gaussian with variance $\sigma^2_{\vartheta,k}$, which is plotted against k in Fig. 1(c).

At each node, the projection order was fixed at $P=4$ and the initial taps were chosen to be zero. The step size μ was set at 0.35 for all the nodes. Similar to [17], the regularization coefficients ρ_{kl} were set to $\rho_{kl}=|\mathcal{N}_k\setminus C(k)|^{-1}$ (the symbol $|\cdot|$ denotes the cardinality of the set) for $l\in\mathcal{N}_k\setminus C(k)$ and $\rho_{kl}=0$ for any other l. Further, $\rho_{kk}=1$ if $\mathcal{N}_k\setminus C(k)=\emptyset$. In most of the literature, the common practice is to choose either the Metropolis rule or the average rule to obtain the combining coefficients a_{lk}. In our simulation studies, we chose the Metropolis rule [1] that gives a doubly stochastic combining matrix. While estimating the error variances in the proposed clustered multitask diffusion APA (i.e., (6)), the mixing coefficient γ was set to 0.5. At first, we consider the

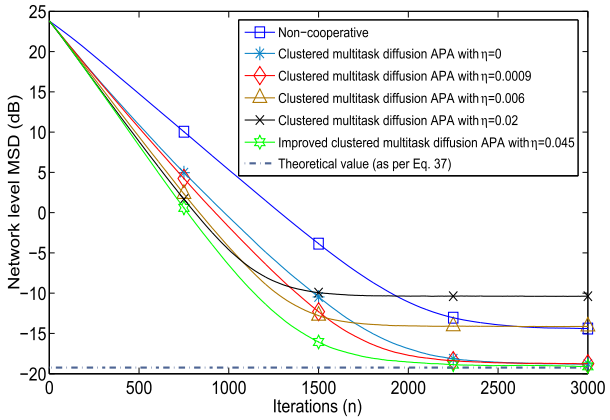


Fig. 2. Network level MSD curves of the proposed improved clustered multitask diffusion APA. Also shown are the network level MSD of non-cooperative APA and conventional clustered multitask diffusion APA for different values of η.

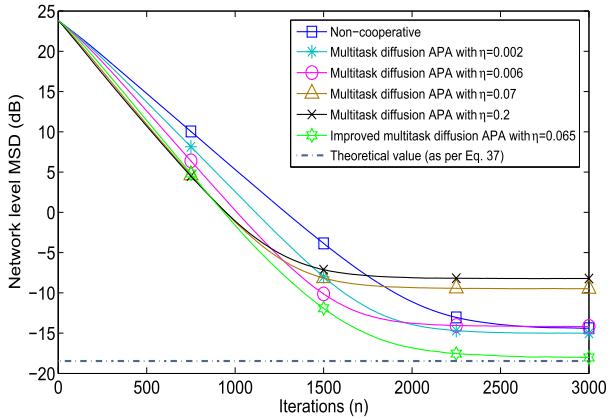


Fig. 3. Network level MSD curves of the proposed improved multitask diffusion APA. Also shown are the network level MSD of non-cooperative APA and conventional multitask diffusion APA for different values of η.

scenario where the cluster information is available. Under this, the proposed improved clustered multitask diffusion strategy was simulated and the simulation results are displayed by plotting the network level MSD (in dB) against the iteration index n, obtained by averaging over 100 independent experiments. The resulting plots are shown in Fig. 2. For comparative assessment, same identification exercise was also carried out by the conventional clustered multitask diffusion APA [22] for different values of regularization strength parameter η, and also by the non-cooperative APA (obtained by setting $\eta = 0$ and the combiner matrix $\mathbf{A} = \mathbf{I}_N$ in the clustered multitask diffusion APA) with other parameters remaining same as used above. The network level MSD curves of these algorithms are also presented in Fig. 2.

Fig. 2 presents an interesting comparison of the proposed improved clustered multitask diffusion APA with the conventional clustered multitask diffusion APA. Firstly, for small values of η (e.g., $0/0.0009$), the conventional clustered multitask algorithm achieves much less steady-state MSD at the cost of a slow convergence rate, but with η increasing to the higher values (e.g., $0.006/0.02$), its convergence becomes faster, but the steady-state MSD increases presumably because of the undesirable effect of inter-cluster cooperation. The improved clustered multitask diffusion APA is, however, seen to enjoy both faster convergence rate and lesser steady-state MSD simultaneously, by exploiting the inter-cluster cooperation in a controlled manner.

Next, we consider the scenario where the cluster information is unavailable. Under this, the improved multitask diffusion strategy (which was obtained by assigning a cluster to each node, i.e., setting $\eta \neq 0$, and the combiner matrix $\mathbf{A} = \mathbf{I}_N$ in the improved clustered multitask diffusion APA) was simulated and the simulation results are plotted in Fig. 3. For comparative assessment, the non-cooperative APA and the conventional multitask diffusion APA [22] were also simulated for different values of η, and the results are plotted in Fig. 3.

From Fig. 3, it can be observed that for small values of η (e.g., $\eta = 0.002$), the conventional multitask diffusion APA exhibits superior performance over non-cooperative strategy in terms of both convergence rate and steady-state MSD. However, as η increases, say to $\eta = 0.006$, the convergence rate increases further with slight degradation in steady-state MSD. For $\eta = 0.006$, the steady-state MSD of the conventional multitask diffusion APA is at par with that for the non-cooperative strategy and the convergence rate is greatly improved. Beyond this point, as η increases further, we observe very little improvement in convergence rate but significant degradation in steady-state MSD performance. Beyond the value of $\eta = 0.07$, it is seen that the steady-state MSD degrades further with no improvement in convergence rate. On the other hand, it can be seen that the improved multitask diffusion APA exhibits superior performance (i.e., simultaneously achieves faster convergence rate and lower steady-state MSD) over the conventional multitask diffusion APA. In fact, the improved multitask diffusion APA is seen to be able to achieve at least 3 dB improvement in the steady-state MSD performance over the conventional multitask diffusion APA which increases further with increase in the value of η. For both the above experiments, in Figs. 2 and 3, we have also plotted the theoretical MSD by horizontal dashed line. The theoretical results show good agreement with the experimental results.

To observe the effectiveness of the proposed improved strategy at node level, we evaluated the steady-state MSD at each individual node and plotted them against the node index in Figs. 4(a) and 4(b) for the clustered, and non-clustered cases, respectively. For comparison, we also plotted the node level steady-state normalized MSD of the conventional strategies in Figs. 4(a) and 4(b). From these figures, it can be observed that the proposed improved strategy performs at par or even better than the non-cooperative strategy at every node which the other schemes fail to achieve.

Finally, to demonstrate the effectiveness of the proposed algorithm in practical applications, we considered the example of distributed acoustic echo cancellation (dAEC). To improve intelligibility of the far-end speech signal, a distributed acoustic echo canceler aims to cancel the acoustic echo (i.e., near-end speech signal) by exploiting the spatial diversity of the acoustic field (i.e., exchanging information among nodes that are monitoring the acoustic field). Since agents (i.e., microphones) are placed at different locations, it is obvious that different acoustic transfer functions exist for each

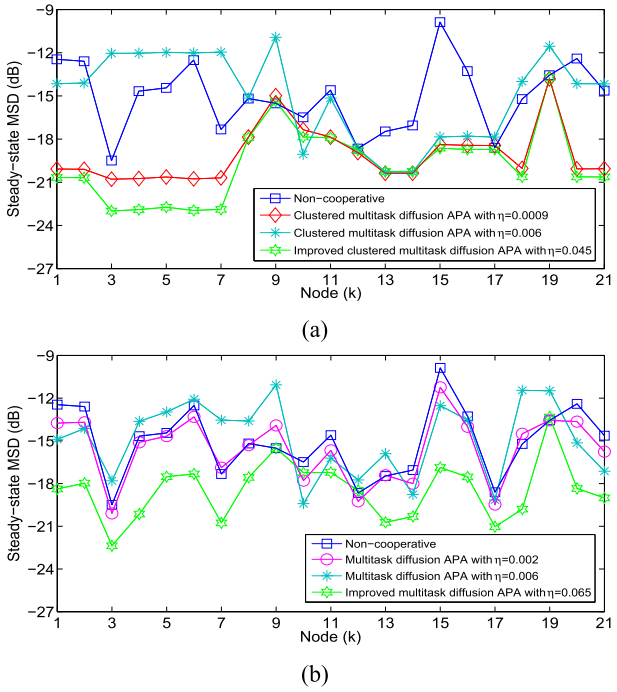


Fig. 4. Node level steady-state MSD of the proposed strategy: (a) improved clustered multitask diffusion APA, (b) improved multitask diffusion APA.

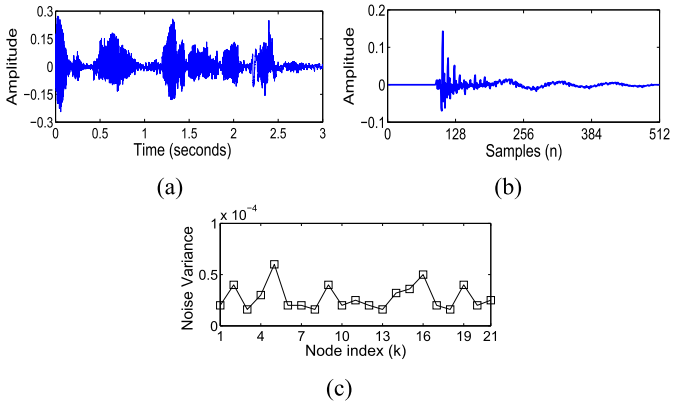


Fig. 5. (a) Far-end speech signal, (b) acoustic echo path $\mathbf{w}_0$, (c) noise variance $\sigma^2_{\vartheta,k}$.

node/agent. Therefore, it is always beneficial to model the problem as a multiple tasks learning problem rather than single task learning. In the following, we evaluate the performance of the proposed algorithm in the context of distributed acoustic echo cancellation.

For this, we considered an acoustic field that was monitored by the above network of 21 interconnected nodes. At each node, the speech signal shown in Fig. 5(a) was used as the far-end signal. For modeling the echo path, the network was divided into four clusters and for each q-th cluster, $q = 1, 2, 3, 4$, the echo path was modeled by a 512×1 coefficient vector of the form $\mathbf{w}^*_{C_q} = \mathbf{w}_0 + \delta_{C_q}\mathbf{w}_q$, for which the echo path impulse response shown in Fig. 5(b) was taken as $\mathbf{w}_0$, and the coefficient δ_{C_q} and the coefficient vector $\mathbf{w}_q$ were generated from zero mean, Gaussian distribution with variance 0.01 and 0.06, respectively. The observation

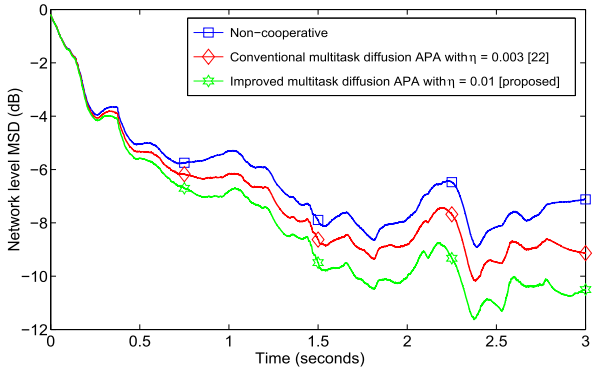


Fig. 6. Learning curve of the proposed improved multitask diffusion APA in distributed acoustic echo cancellation setting. Also shown are the learning curves of the non-cooperative APA and the conventional multitask diffusion APA.

noise $\vartheta_k(n)$ at the k-th node was taken to be zero mean i.i.d. Gaussian with variance $\sigma^2_{\vartheta,k}$, which is plotted against k in Fig. 5(c). For assessing the performance of the proposed algorithm, we, however, considered the worst case scenario where no cluster information was assumed to be available and the pure multitask diffusion APA (i.e., where each node is a cluster) with $P = 4$, along with the proposed controlled inter-node collaboration, was used to cancel the near-end signal. The corresponding learning curve (MSD vs time index) is displayed in Fig. 6. For comparison, we also plotted the learning curves of the conventional multitask diffusion APA and the non-cooperative APA. From Fig. 6, it shall be observed that the proposed controlled inter-cluster cooperation helps to cancel the near-end signal more effectively than the other two methods by restricting the inter-cluster cooperation in the right direction.

V. Conclusion

An APA based multitasking, distributed adaptive filter is proposed for multiple parameter vector estimation under temporarily correlated input conditions. The proposed algorithm achieves remarkably improved performance over state-of-the-art in terms of both convergence rate and steady-state MSD, by deploying a controlled form of inter-cluster collaboration via a control variable which enables the network to maintain collaboration in right direction. A detailed performance analysis of the proposed algorithm is carried out and stability bounds are obtained for both mean and mean square convergence. The claims made are validated via exhaustive simulation studies.

Appendix A
Proof of Theorem 1

Taking expectations on both sides of (18) and using the assumptions 2 and 4, we obtain

$$E\big[\widetilde{\mathbf{w}}(n+1)\big] = E[\boldsymbol{\mathcal{B}}_\delta(n)]E[\widetilde{\mathbf{w}}(n)] + E[\mathbf{r}_\delta(n)], \qquad (39)$$

where $E[\boldsymbol{\mathcal{B}}_\delta(n)] = \boldsymbol{\mathcal{A}}\big(\mathbf{I}_{LN} - \mu\overline{\mathbf{Z}} - \mu\eta E\big[\boldsymbol{\mathcal{Q}}_\delta(n)\big]\big)$ and $E[\mathbf{r}_\delta(n)] = \mu\eta\boldsymbol{\mathcal{A}}E\big[\boldsymbol{\mathcal{Q}}_\delta(n)\big]\mathbf{w}^\star$. The matrix $E[\boldsymbol{\mathcal{Q}}_\delta(n)]$ is given by

$$E[\boldsymbol{\mathcal{Q}}_\delta(n)] = \big(E\big[\mathbf{D}_\delta(n)\big] \otimes \mathbf{I}_L\big) - \big(E\big[\mathbf{P}_\delta(n)\big] \otimes \mathbf{I}_L\big), \qquad (40)$$

with

$$E[\mathbf{D}_\delta(n)] = \text{diag}\{E[\rho_{\delta_1}(n)], \cdots, E[\rho_{\delta_N}(n)]\},$$
$$E[\mathbf{P}_{\boldsymbol{\delta}}(n)] = \mathbf{P} \odot E[\boldsymbol{\delta}(n)], \tag{41}$$

and

$$E[\rho_{\delta_k}(n)] = \sum_{l \in \mathcal{N}_k \backslash C(k)} E[\rho_{\delta_{kl}}(n)] = \sum_{l \in \mathcal{N}_k \backslash C(k)} \rho_{kl} E[\delta_{kl}(n)] \quad \text{for} \quad k = 1, 2, \cdots, N. \tag{42}$$

Iterating the recursion (39), backwards down to $n = 0$,

$$E\big[\widetilde{\mathbf{w}}(n)\big] = \Big(\prod_{i=0}^{n-1} E[\mathcal{B}_\delta(i)]\Big) E[\widetilde{\mathbf{w}}(0)] + \sum_{i=0}^{n-2} \Big(\prod_{j=i+1}^{n-1} E[\mathcal{B}_\delta(j)]\Big) E[\mathbf{r}_\delta(i)] + E[\mathbf{r}_\delta(n-1)]. \tag{43}$$

A sufficient condition for $\lim_{n\to\infty} E\big[\widetilde{\mathbf{w}}(n)\big]$ to attain a finite value is that $\|E[\mathcal{B}_\delta(n)]\| < 1$ for all n, where $\|\cdot\|$ is any matrix norm. To derive a convergence condition in terms of μ, we use the block maximum norm of the matrix $E[\mathcal{B}_\delta(n)]$ (i.e., $\|E[\mathcal{B}_\delta(n)]\|_{b,\infty}$). From the properties of the block maximum norm [1], we can write

$$\begin{aligned}\|E[\mathcal{B}_\delta(n)]\|_{b,\infty} &= \big\|\mathcal{A}\big(\mathbf{I}_{LN} - \mu\overline{\mathbf{Z}} - \mu\eta E[\mathcal{Q}_\delta(n)]\big)\big\|_{b,\infty} \\ &\le \|\mathcal{A}\|_{b,\infty} \big\|\mathbf{I}_{LN} - \mu\overline{\mathbf{Z}} - \mu\eta E[\mathcal{Q}_\delta(n)]\big\|_{b,\infty} \\ &= \big\|\mathbf{I}_{LN} - \mu\overline{\mathbf{Z}} - \mu\eta E[\mathcal{Q}_\delta(n)]\big\|_{b,\infty}. \end{aligned} \tag{44}$$

In the above, we used the result $\|\mathcal{A}\|_{b,\infty} = \|\mathbf{A}^T\|_\infty = 1$. Substituting (40) in (44) and using the block maximum norm properties, we get

$$\begin{aligned}\|E[\mathcal{B}_\delta(n)]\|_{b,\infty} &\le \|\mathbf{I}_{LN} - \mu\overline{\mathbf{Z}} - \mu\eta\big(E\big[\mathbf{D}_\delta(n)\big] \otimes \mathbf{I}_L\big)\|_{b,\infty} \\ &\quad + \mu\eta\|E[\mathbf{P}_\delta(n)] \otimes \mathbf{I}_L\|_{b,\infty} \\ &= \rho\big(\mathbf{I}_{LN} - \mu\overline{\mathbf{Z}} - \mu\eta(E[\mathbf{D}_\delta(n)] \otimes \mathbf{I}_L)\big) \\ &\quad + \mu\eta\|E[\mathbf{P}_\delta(n)]\|_\infty \\ &= \rho\big(\mathbf{I}_{LN} - \mu\overline{\mathbf{Z}} - \mu\eta(E[\mathbf{D}_\delta(n)] \otimes \mathbf{I}_L)\big) \\ &\quad + \mu\eta \max_{1\le k\le N}\{E[\rho_{\delta_k}(n)]\}. \end{aligned} \tag{45}$$

From these, a sufficient condition for $E[\widetilde{\mathbf{w}}(n)]$ to converge is $\rho\big(\mathbf{I}_{LN} - \mu\overline{\mathbf{Z}} - \mu\eta(E[\mathbf{D}_\delta(n)] \otimes \mathbf{I}_L)\big) + \mu\eta \max_{1\le k\le N}\{E[\rho_{\delta_k}(n)]\} < 1$, or, $-1 + \mu\eta \max_{1\le k\le N}\{E[\rho_{\delta_k}(n)]\} < 1 - \mu\lambda_i(\overline{\mathbf{Z}}) - \mu\eta\lambda_j(E[\mathbf{D}_\delta(n)]) < 1 - \mu\eta \max_{1\le k\le N}\{E[\rho_{\delta_k}(n)]\}$, $i = 1, 2, \cdots, LN$ and $j = 1, 2, \cdots, N$ (where $i = (j-1)L + r$), $r = 1, 2, \cdots, L$); which leads to $0 < \mu < \frac{2}{\lambda_i(\overline{\mathbf{Z}}) + \eta\lambda_j(E[\mathbf{D}_\delta(n)]) + \eta \max_{1\le k\le N}\{E[\rho_{\delta_k}(n)]\}}$. Thus, a sufficient condition for convergence is given by $0 < \mu < \frac{2}{\max_{i=1,\cdots,LN} \lambda_i(\overline{\mathbf{Z}}) + 2\eta \max_{1\le k\le N}\{E[\rho_{\delta_k}(n)]\}}$, which proves (19).

APPENDIX B
PROOF OF THEOREM 2

Iterating the recursion (34) backwards down to $n = 0$, we get

$$\begin{aligned} &E[\|\widetilde{\mathbf{w}}(n+1)\|^2_{\text{bvec}^{-1}\{\boldsymbol{\sigma}\}}] \\ &= E[\|\widetilde{\mathbf{w}}(0)\|^2_{\text{bvec}^{-1}\{\left(\prod_{i=0}^{n} \boldsymbol{\mathcal{F}}_\delta^T(i)\right)\boldsymbol{\sigma}\}}] \\ &\quad + \mu^2 \boldsymbol{\gamma}^T\big(\mathbf{I}_{L^2N^2} + \sum_{i=1}^{n}\big(\prod_{j=i}^{n} \boldsymbol{\mathcal{F}}_\delta^T(j)\big)\big)\boldsymbol{\sigma} \\ &\quad + \sum_{i=0}^{n-1} \mathbf{f}\left(\begin{matrix}\mathbf{r}_{b,\delta}(n-1-i), \boldsymbol{\alpha}_\delta(n-1-i), \\ E[\widetilde{\mathbf{w}}(n-1-i)], \big(\prod_{j=0}^{i} \boldsymbol{\mathcal{F}}_\delta^T(n-j)\big)\boldsymbol{\sigma}\end{matrix}\right) \\ &\quad + \mathbf{f}\big(\mathbf{r}_{b,\delta}(n), \boldsymbol{\alpha}_\delta(n), E[\widetilde{\mathbf{w}}(n)], \boldsymbol{\sigma}\big) \end{aligned} \tag{46}$$

where $\widetilde{\mathbf{w}}(0) = \mathbf{w}^* - \mathbf{w}(0)$. Note that under $\|\boldsymbol{\mathcal{F}}_\delta(n)\| < 1$, $\big(\prod_{i=0}^{n} \boldsymbol{\mathcal{F}}_\delta^T(i)\big)\boldsymbol{\sigma} \to 0$ and $\sum_{i=1}^{n}\big(\prod_{j=i}^{n} \boldsymbol{\mathcal{F}}_\delta^T(j)\big)\boldsymbol{\sigma} \to$ a finite quantity as $n \to \infty$ ($\|\cdot\|$ denotes any matrix norm). A sufficient condition for convergence of $E[\|\widetilde{\mathbf{w}}(n+1)\|^2_{\text{bvec}^{-1}\{\boldsymbol{\sigma}\}}]$ is then given by $\|\boldsymbol{\mathcal{F}}_\delta(n)\| < 1$. To derive a convergence condition in terms of μ, we use the block maximum norm of the matrix $\boldsymbol{\mathcal{F}}_\delta(n)$ (i.e., $\|\boldsymbol{\mathcal{F}}_\delta(n)\|_{b,\infty}$).

From the properties of block maximum norm, we can write

$$\begin{aligned}\|\boldsymbol{\mathcal{F}}_\delta(n)\|_{b,\infty} &= \big\|\big(\boldsymbol{\mathcal{A}} \otimes_b \boldsymbol{\mathcal{A}}\big)\boldsymbol{\mathcal{H}}_\delta(n)\big\|_{b,\infty} \\ &\le \big\|\big(\boldsymbol{\mathcal{A}} \otimes_b \boldsymbol{\mathcal{A}}\big)\big\|_{b,\infty} \big\|\boldsymbol{\mathcal{H}}_\delta(n)\big\|_{b,\infty}. \end{aligned} \tag{47}$$

Since $\mathbf{A}$ is a left stochastic matrix and $\boldsymbol{\mathcal{A}} \otimes_b \boldsymbol{\mathcal{A}} = (\mathbf{A} \otimes \mathbf{A})^T \otimes (\mathbf{I}_L \otimes \mathbf{I}_L)$, from properties of block maximum norm, we have $\|\boldsymbol{\mathcal{A}} \otimes_b \boldsymbol{\mathcal{A}}\|_{b,\infty} = \|(\mathbf{A} \otimes \mathbf{A})^T \otimes \mathbf{I}_L\|_\infty = 1$. and substituting from the definition of $\boldsymbol{\mathcal{H}}_\delta(n)$ as given by (26) where $E\big[\boldsymbol{\mathcal{Q}}_\delta(n)\big]$ is given by (40), we have

$$\begin{aligned}\|\boldsymbol{\mathcal{F}}_\delta(n)\|_{b,\infty} \le &\left\| \begin{matrix} \mathbf{I}_{L^2N^2} - \mu\big(\mathbf{I}_{LN} \otimes_b \overline{\mathbf{Z}}\big) - \mu\big(\overline{\mathbf{Z}} \otimes_b \mathbf{I}_{LN}\big) \\ -\mu\eta\Big(\begin{matrix}(E[\mathbf{D}_\delta(n)] \otimes \mathbf{I}_L) \otimes_b \mathbf{I}_{LN} \\ +\mathbf{I}_{LN} \otimes_b (E[\mathbf{D}_\delta(n)] \otimes \mathbf{I}_L)\end{matrix}\Big) \end{matrix} \right\|_{b,\infty} \\ &+ \mu\eta\|(E[\mathbf{P}_\delta(n)] \otimes \mathbf{I}_N) \otimes \mathbf{I}_{L^2}\|_{b,\infty} \\ &+ \mu\eta\|(\mathbf{I}_N \otimes E[\mathbf{P}_\delta(n)]) \otimes \mathbf{I}_{L^2}\|_{b,\infty}. \end{aligned} \tag{48}$$

From the properties of block maximum norm, we have $\big\|\big(E[\mathbf{P}_\delta(n)] \otimes \mathbf{I}_N\big) \otimes \mathbf{I}_{L^2}\big\|_{b,\infty} + \big\|\big(\mathbf{I}_N \otimes E[\mathbf{P}_\delta(n)]\big) \otimes \mathbf{I}_{L^2}\big\|_{b,\infty} = \big\|\big(E[\mathbf{P}_\delta(n)] \otimes \mathbf{I}_N\big) \otimes \mathbf{I}_L\big\|_\infty + \big\|\big(\mathbf{I}_N \otimes E[\mathbf{P}_\delta(n)]\big) \otimes \mathbf{I}_L\big\|_\infty = 2 \max_{1\le k\le N}\big\{E[\rho_{\delta_k}(n)]\big\}$. Moreover, as the first term in the R.H.S of (48) is a block diagonal symmetric matrix, its block maximum norm is equal to its spectral radius. Substituting these results in (48), it is seen that $E[\|\widetilde{\mathbf{w}}(n+1)\|^2_{\boldsymbol{\Sigma}}]$ converges under $\rho\big(\mathbf{I}_{L^2N^2} - \mu(\mathbf{I}_{LN} \otimes_b \overline{\mathbf{Z}}) - \mu(\overline{\mathbf{Z}} \otimes_b \mathbf{I}_{LN}) - \mu\eta((E[\mathbf{D}_\delta(n)] \otimes \mathbf{I}_L) \otimes_b \mathbf{I}_{LN} + \mathbf{I}_{LN} \otimes_b (E[\mathbf{D}_\delta(n)] \otimes \mathbf{I}_L))\big) + 2\mu\eta \max_{1\le k\le N}\{E[\rho_{\delta_k}(n)]\} < 1$. First note that for this, we must have $2\mu\eta \max_{1\le k\le N}\{E[\rho_{\delta_k}(n)]\} < 1$. Assuming this to be true, we then make note of the following: since every eigenvector of $\overline{\mathbf{Z}}$, say $\mathbf{e}_i$, $i = 1, 2, \cdots, LN$ with corresponding eigenvalue $\lambda_i(\overline{\mathbf{Z}})$, is also an eigenvector of $\mathbf{I}_{LN}$ (with eigenvalue $=1$), from (17g), it is seen that both $(\mathbf{I}_{LN} \otimes_b \overline{\mathbf{Z}})$ and $(\overline{\mathbf{Z}} \otimes_b \mathbf{I}_{LN})$ have the same set of eigenvectors, namely, $\mathbf{e}_i \otimes_b \mathbf{e}_j$, $i, j = 1, 2, \cdots, LN$ and the same set of eigenvalues, namely, $\lambda_l(\overline{\mathbf{Z}})$, $l = 1, 2, \cdots, LN$ (each $\lambda_l(\overline{\mathbf{Z}})$ has multiplicity LN; also, the eigenvector $\mathbf{e}_i \otimes_b \mathbf{e}_j$ has eigenvalue $\lambda_j(\overline{\mathbf{Z}})$ for $(\mathbf{I}_{LN} \otimes_b \overline{\mathbf{Z}})$ and $\lambda_i(\overline{\mathbf{Z}})$ for $(\overline{\mathbf{Z}} \otimes_b \mathbf{I}_{LN})$). Therefore, the above convergence condition can be equivalently stated as $|1 - \mu(\lambda_i(\overline{\mathbf{Z}}) + \lambda_j(\overline{\mathbf{Z}})) - \mu\eta(\lambda_r(E[\mathbf{D}_\delta(n)]) + \lambda_s(E[\mathbf{D}_\delta(n)]))| < 1 - 2\mu\eta \max_{1\le k\le N}\{E[\rho_{\delta_k}(n)]\}$, $i, j = 1, 2, \cdots, LN$ and $r, s = 1, 2, \cdots, N$; which leads to $0 < \mu < \frac{2}{\lambda_i(\overline{\mathbf{Z}}) + \lambda_j(\overline{\mathbf{Z}}) + \eta(\lambda_r(E[\mathbf{D}_\delta(n)]) + \lambda_s(E[\mathbf{D}_\delta(n)])) + 2\eta \max_{1\le k\le N}\{E[\rho_{\delta_k}(n)]\}}$.

Thus, a sufficient condition for convergence is given by $0 < \mu < \frac{2}{2\max\limits_{i=1,\cdots,LN}\lambda_i(\overline{\mathbf{Z}})+4\eta\max\limits_{1\le k\le N}\{E[\rho_{\delta_k}(n)]\}}$, which proves (35).

REFERENCES

[1] A. H. Sayed, "Diffusion adaptation over networks," in *Academic Press Library in Signal Processing*, R. Chellappa and S. Theodoridis, Eds. Amsterdam, The Netherlands: Elsevier, 2013, pp. 322–454.

[2] A. H. Sayed, "Adaptive networks," *Proc. IEEE*, vol. 102, no. 4, pp. 460–497, Apr. 2014.

[3] A. H. Sayed, *Adaptation, Learning, and Optimization Over Networks*. Boston, MA, USA: Now, 2014.

[4] C. G. Lopes and A. H. Sayed, "Incremental adaptive strategies over distributed networks," *IEEE Trans. Signal Process.*, vol. 55, no. 8, pp. 4064–4077, Aug. 2007.

[5] L. Li, J. A. Chambers, C. G. Lopes, and A. H. Sayed, "Distributed estimation over an adaptive incremental network based on the affine projection algorithm," *IEEE Trans. Signal Process.*, vol. 58, no. 1, pp. 151–164, Jan. 2010.

[6] S. Kar and J. M. F. Moura, "Distributed consensus algorithms in sensor networks with imperfect communication: Link failures and channel noise," *IEEE Trans. Signal Process.*, vol. 57, no. 1, pp. 355–369, Jan. 2009.

[7] J. Zhan and X. Li, "Cluster consensus in networks of agents with weighted cooperative-competitive interactions," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 65, no. 2, pp. 241–245, Feb. 2018.

[8] F. S. Cattivelli and A. H. Sayed, "Diffusion LMS strategies for distributed estimation," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1035–1048, Mar. 2010.

[9] S. Zhang, H. C. So, W. Mi, and H. Han, "A family of adaptive decorrelation NLMS algorithms and its diffusion version over adaptive networks," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 65, no. 2, pp. 638–649, Feb. 2018.

[10] A. Rastegarnia, "Reduced-communication diffusion RLS for distributed estimation over multi-agent networks," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, to be published, doi: 10.1109/TCSII.2019.2899194.

[11] L. Li and J. A. Chambers, "Distributed adaptive estimation based on the APA algorithm over diffusion networks with changing topology," in *Proc. IEEE/SP Workshop Stat. Signal Process.*, Cardiff, U.K., 2009, pp. 757–760.

[12] S.-Y. Tu and A. H. Sayed, "Diffusion strategies outperform consensus strategies for distributed estimation over adaptive networks," *IEEE Trans. Signal Process.*, vol. 60, no. 12, pp. 6217–6234, Dec. 2012.

[13] J. Plata-Chaves, A. Bertrand, and M. Moonen, "Incremental multiple error filtered-X LMS for node-specific active noise control over wireless acoustic sensor networks," in *Proc. IEEE Sensor Array Multichannel Signal Process. Workshop (SAM)*, Rio de Janerio, Brazil, Jul. 2016, pp. 1–5.

[14] J. Plata-Chaves, A. Bertrand, M. Moonen, S. Theodoridis, and A. M. Zoubir, "Heterogeneous and multitask wireless sensor networks-algorithms, applications, and challenges," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 3, pp. 450–465, Apr. 2017.

[15] A. Hassani, J. Plata-Chaves, M. H. Bahari, M. Moonen, and A. Bertrand, "Multi-task wireless sensor network for joint distributed node-specific signal enhancement, LCMV beamforming and DOA estimation," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 3, pp. 518–533, Mar. 2017.

[16] J. Chen, C. Richard, and A. H. Sayed, "Diffusion LMS for clustered multitask networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Florence, Italy, 2014, pp. 5487–5491.

[17] J. Chen, C. Richard, and A. H. Sayed, "Multitask diffusion adaptation over networks," *IEEE Trans. Signal Process.*, vol. 62, no. 16, pp. 4129–4144, Aug. 2014.

[18] J. Plata-Chaves, N. Bogdanovi, and K. Berberidis, "Distributed diffusion-based LMS for node-specific adaptive parameter estimation," *IEEE Trans. Signal Process.*, vol. 63, no. 13, pp. 3448–3460, Jul. 2015.

[19] R. Nassif, C. Richard, A. Ferrari, and A. H. Sayed, "Multitask diffusion adaptation over asynchronous networks," *IEEE Trans. Signal Process.*, vol. 64, no. 11, pp. 2835–2850, Jun. 2016.

[20] M. Hajiabadi, G. A. Hodtani, and H. Khoshbin, "Robust learning over multitask adaptive networks with wireless communication links," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 66, no. 6, pp. 1083–1087, Jun. 2019.

[21] K. Ozeki and T. Umeda, "An adaptive filtering algorithm using an orthogonal projection to an affine subspace and its properties," *Electron. Commun. Jpn.*, vol. 67-A, no. 5, pp. 19–27, 1984.

[22] V. C. Gogineni and M. Chakraborty, "Diffusion affine projection algorithm for multitask networks," in *Proc. IEEE Int. Conf. Asia–Pacific Signal Inf. Process. Assoc. (APSIPA)*, Honolulu, HI, USA, Nov. 2018, pp. 201–206.

[23] H.-C. Shin and A. H. Sayed, "Mean-square performance of a family of affine projection algorithms," *IEEE Trans. Signal Process.*, vol. 52, no. 1, pp. 90–102, Jan. 2004.

[24] R. L. Das and M. Chakraborty, "Sparse adaptive filters—An overview and some new results," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Seoul, South Korea, May 2012, pp. 2745–2748.

[25] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

Vinay Chakravarthi Gogineni received the bachelor's degree in electronics and communication engineering from Jawaharlal Nehru Technological University, India, in 2005, the master's degree in communication engineering from the Vellore Institute of Technology, Vellore, India, in 2008, and the Ph.D. degree in electronics and electrical communication engineering from the Indian Institute of Technology Kharagpur, India, in 2019.

From 2008 to 2011, he was with Cosyres Technologies, Bengaluru, India, where he was involved in algorithmic-level optimization of audio codecs. He is currently an European Research Consortium for Informatics and Mathematics (ERCIM) Postdoctoral Research Fellow with the Department of Electronic Systems, NTNU, Norway. His research interests include adaptive filtering and machine learning, statistical signal processing, and graph signal processing.

Mrityunjoy Chakraborty (M'94–SM'99) is currently a Professor in electronics and electrical communication engineering with the Indian Institute of Technology, Kharagpur. His teaching and research interests are digital and adaptive signal processing, VLSI DSP, linear algebra and compressive sensing.

Prof. Chakraborty has been a member of the APSIPA BOG from 2013 to 2016. He is a fellow of the National Academy of Sciences, India, and the Indian National Academy of Engineering (INAE). He is currently a Senior Editorial Board (SEB) Member of the *IEEE Signal Processing Magazine* and also served the IEEE *Journal on Emerging Techniques in Circuits and Systems* as an SEB Member from 2016 to 2017. He received the prestigious Chair Professorship of the INAE. He has been the DSP Track Co-Chair of ISCAS from 2015 to 2020, the TPC Co-Chair of the IEEE SIPS-2018, the special Session Co-Chair of DSP-18, and the Gabor Track Chair of DSP-15. He has been the General Chair of the National Conference on Communications in 2012 and 2020. He served as a Chair of the DSP Technical Committee (TC) of the IEEE Circuits and Systems Society from 2016 to 2018 and the APSIPA TC on Signal and Information Processing Theory and Methods. He has been a Guest Editor of the EURASIP JASP and special issues of TCAS-II. Earlier, he had been an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I: REGULAR PAPERS from 2004 to 2007 and from 2010 to 2012, and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—II: EXPRESS BRIEFS from 2008 to 2009. From 2012 to 2013, he was selected as a Distinguished Lecturer of the APSIPA. He is a Co-Founder of the APSIPA.