TRANSIENT ANALYSIS OF CLUSTERED MULTITASK DIFFUSION RLS ALGORITHM

Wei Gao † *Jie Chen* ⋆ *Cédric Richard* ‡ *Wentao Shi* ⋆ *Qunfei Zhang* ⋆

†School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang 212013, China
⋆School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China
‡Université Côte d'Azur, OCA, CNRS, 06108 Nice, France
Email: wei_gao@ujs.edu.cn dr.jie.chen@ieee.org cedric.richard@unice.fr swt@nwpu.edu.cn zhangqf@nwpu.edu.cn

ABSTRACT

In this paper, we propose a novel clustered multitask diffusion RLS (MT-DRLS) algorithm over network to further improve the performance of its counterpart, the multitask diffusion LMS (MT-DLMS) algorithm. Its transient behavior is investigated, in the mean and mean-square error sense. Simulation results illustrate the significant improvement of the MT-DRLS over the MT-DLMS in terms of convergence rate and steady-state error, as well as the accuracy of the theoretical findings.

Index Terms— Multitask adaptive learning, diffusion RLS, transient analysis, distributed estimation, multitask network.

1. INTRODUCTION

During the past decade, distributed detection, estimation, and tracking problems have attracted substantial attention in the context of adaptive networks with diffusion strategies [1–3]. Particularly, it has been found that multitask networks have a wider range of modeling ability than single-task networks. Several multitask strategies for adaptation and learning over networks have been recently proposed based on the diffusion least-mean-squares (DLMS) algorithm [4–11]. To be specific, the DLMS algorithm for multitask networks was first proposed in [4, 5], and studied in asynchronous networks in [6]. A new multitask learning formulation using a common latent representation was presented in [7], as well as a unified framework to analyze its performance. Both ℓ_1-norm regularization and $\ell_{\infty,1}$-norm regularization were introduced into multitask networks in [9] and [10], respectively. Recently, the performance of multitask DLMS algorithm has been analyzed in the presence of communication delays [11]. An overview of multitask learning over networks and its applications is available in [12].

The diffusion recursive least-squares (DRLS) algorithm was also extensively studied in [13–20], due to the superior performance of the RLS compared to the LMS [21, 22]. The DRLS algorithm with incremental update-then-combine diffusion strategy was initially proposed in [13], with an analysis of its steady-state performance. The diffusion bias-compensated RLS algorithm was presented in [14] to reduce residual bias, and the DRLS algorithm was considered in [15, 19] to reduce communication cost. Variants of the DRLS algorithm were successively devised to improve the performance in the context of sparse systems [16], noisy links [17], and robustness against impulsive interferences [18]. More recently, a transient analysis of DRLS algorithm was presented in [20]. To the best of our knowledge, the multitask DRLS algorithm has not been considered so far except in [23]. This motivates us to derive in this paper the clustered multitask diffusion RLS (MT-DRLS) algorithm with adapt-then-combine (ATC) diffusion strategy. Furthermore, analytical models are derived to characterize its transient behavior in the mean and mean-square error sense. Simulation results illustrate the superiority of the MT-DRLS algorithm over the MT-DLMS and DRLS algorithms. The accuracy of the resulting transient analytical models is also investigated.

Notation: The matrix trace is denoted by $\mathrm{tr}\{\cdot\}$. The notation $\otimes$ denotes Kronecker product. Identity matrix of size $N\times N$ is denoted by $\mathbf{I}_N$, and $\mathbf{1}_N$ denotes an all-one vector of length N. The operator $\mathrm{bdiag}\{\cdot\}$ formulates a (block) diagonal matrix with its arguments, and $\mathrm{col}\{\cdot\}$ stacks its vector arguments on the top of each other to generate a column vector. The notation $\|\mathbf{x}\|^2_{\mathbf{\Sigma}}$ denotes the squared norm of $\mathbf{x}$ weighted by any positive semi-definite matrix $\mathbf{\Sigma}$, i.e., $\|\mathbf{x}\|^2_{\mathbf{\Sigma}}=\mathbf{x}^\top\mathbf{\Sigma}\mathbf{x}$. $\mathcal{N}_k$ denotes the neighborhood of node k, including k. $\mathcal{C}(k)$ and $\mathcal{C}(k)^-$ denote the cluster of nodes to which node k belongs, including k and excluding k, respectively.

2. NETWORK MODEL AND MT-DRLS ALGORITHM

2.1. Clustered Multitask Network Model

Consider a connected network consisting of K nodes, indexed with $k=1,\ldots,K$. At time instant $n\geq 0$, each node k has access to a random data pair $\{d_{k,n},\mathbf{x}_{k,n}\}$, which is assumed to be generated by an optimal weight vector $\mathbf{w}_k^\star\in\mathbb{R}^L$ at node k via the linear regression model:

$$d_{k,n}=\mathbf{x}_{k,n}^\top\mathbf{w}_k^\star+z_{k,n} \tag{1}$$

where $d_{k,n}\in\mathbb{R}$ is the zero-mean desired signal, $\mathbf{x}_{k,n}\in\mathbb{R}^L$ denotes the regression vector with a positive-definite covariance matrix $\mathbf{R}_{x,k}=\mathbb{E}\{\mathbf{x}_{k,n}\mathbf{x}_{k,n}^\top\}$, and $z_{k,n}$ is a zero-mean temporally and spatially independent noise with variance $\sigma^2_{z,k}$. We assume that all nodes are grouped into Q clusters, corresponding to Q estimation tasks. The unknown optimal weight vector $\mathbf{w}_k^\star$ are constrained to be identical within each cluster, namely, $\mathbf{w}_k^\star=\mathbf{w}_{\mathcal{C}_q}^\star$ for all $k\in\mathcal{C}_q$, where $\mathcal{C}_q$ denotes the cluster q, i.e., the index set of nodes in the q-th cluster. However, it is assumed that similarities exist among the neighboring clusters, i.e., $\mathbf{w}_{\mathcal{C}_p}^\star\sim\mathbf{w}_{\mathcal{C}_q}^\star$ if $\mathcal{C}_p$ and $\mathcal{C}_q$ are connected with $p\neq q$, where $\sim$ represents a similarity relationship in some sense. Clusters $\mathcal{C}_p$ and $\mathcal{C}_q$ are connected provided that there exists at least one communication link connecting a node from one cluster to a node in the other cluster.

2.2. Clustered Multitask Diffusion RLS Algorithm

In the context of clustered multitask networks, the objective is to estimate the unknown parameter vectors $\{\mathbf{w}_{\mathcal{C}_q}^\star\}_{q=1}^Q$. For each node k

This work was supported in part by the National Natural Science Foundation of China Grants (62171205, 62171380).

in cluster$\mathcal{C}(k)$ and nodes of other clusters connecting to node k, we first introduce an intermediate similarity promotion equation to promote the similarities of weight vectors between neighboring clusters:

$$\mathbf{w}_{k,n-\frac{1}{2}} = \mathbf{w}_{k,n-1} - \gamma \sum_{\ell \in \mathcal{N}_k \setminus \mathcal{C}(k)^-} \frac{\rho_{k\ell} + \rho_{\ell k}}{2} (\mathbf{w}_{k,n-1} - \mathbf{w}_{\ell,n-1}) \tag{2}$$

with the set difference $\setminus$ and the strength parameter $\gamma \geq 0$, and where $\mathbf{w}_{k,n-1}$ denotes the local estimate of $\mathbf{w}_k^\star$. Here, the non-negative weight coefficients $\rho_{k\ell}$ are chosen to satisfy the conditions [5, 6]:

$$\sum_{\ell \in \mathcal{N}_k \setminus \mathcal{C}(k)^-} \rho_{k\ell} = 1, \text{ and } \begin{cases} \rho_{k\ell} > 0, & \text{if } \ell \in \mathcal{N}_k \setminus \mathcal{C}(k), \\ \rho_{kk} \geq 0, & \\ \rho_{k\ell} = 0, & \text{otherwise.} \end{cases} \tag{3}$$

We collect the above coefficients into the random $K \times K$ right-stochastic matrix $\boldsymbol{\Theta}$ with (k, ℓ)-th entry $\rho_{k\ell}$. Note that the weight coefficients between pairs of nodes can be assumed to be symmetric or asymmetric. We use the symmetric case here for simplicity.

Let $\boldsymbol{\psi}_{k,n}$ denote the intermediate estimate of $\mathbf{w}_k^\star$. Since the weight vector $\mathbf{w}_{k,n-\frac{1}{2}}$ that considers the similarities of weight vectors between neighboring clusters is a good guess for the intermediate estimate $\boldsymbol{\psi}_{k,n}$ as a prior information, we then consider a special case of local least-squares problem based on the collected data at node k only for current time instant n as in [18, 20]:

$$\boldsymbol{\psi}_{k,n} = \underset{\boldsymbol{\psi}_k \in \mathbb{R}^L}{\arg\min} \left\{ \left\| \boldsymbol{\psi}_k - \mathbf{w}_{k,n-\frac{1}{2}} \right\|_{\boldsymbol{\Lambda}_{k,n}}^2 + \left(d_{k,n} - \mathbf{x}_{k,n}^\top \boldsymbol{\psi}_k \right)^2 \right\} \tag{4}$$

with the positive-definite weighting matrix $\boldsymbol{\Lambda}_{k,n}$ and the time-averaged autocorrelation matrix $\boldsymbol{\Phi}_{k,n}$ of input data for node k at time instant n, namely,

$$\boldsymbol{\Lambda}_{k,n} = \boldsymbol{\Phi}_{k,n} - \mathbf{x}_{k,n} \mathbf{x}_{k,n}^\top, \tag{5}$$

$$\boldsymbol{\Phi}_{k,n} = \lambda \boldsymbol{\Phi}_{k,n-1} + \mathbf{x}_{k,n} \mathbf{x}_{k,n}^\top, \tag{6}$$

where $0 \ll \lambda < 1$ is the forgetting factor, and the initial condition is $\boldsymbol{\Phi}_{k,0} = \delta \mathbf{I}_L$ with a small positive value δ. Considering variable substitutions $\mathbf{v} = \boldsymbol{\psi}_k - \mathbf{w}_{k,n-\frac{1}{2}}$ and $b = d_{k,n} - \mathbf{x}_{k,n}^\top \mathbf{w}_{k,n-\frac{1}{2}}$, problem (4) can be reformulated as:

$$\mathbf{v}^\star = \underset{\mathbf{v} \in \mathbb{R}^L}{\arg\min} \left\{ \mathbf{v}^\top \boldsymbol{\Lambda}_{k,n} \mathbf{v} + \left(b - \mathbf{x}_{k,n}^\top \mathbf{v} \right)^2 \right\}. \tag{7}$$

Setting the derivative of (7) with respect to $\mathbf{v}$ to zero, we obtain:

$$\boldsymbol{\Phi}_{k,n} \mathbf{v} = b\, \mathbf{x}_{k,n}. \tag{8}$$

By the definition of $\mathbf{v}$, applying the matrix inversion lemma to the right hand side (r.h.s.) of (6), then (8) can be rewritten as:

$$\boldsymbol{\psi}_{k,n} = \mathbf{w}_{k,n-\frac{1}{2}} + b\, \mathbf{P}_{k,n} \mathbf{x}_{k,n} \tag{9}$$

with the definition $\boldsymbol{\Phi}_{k,n} = \mathbf{P}_{k,n}^{-1}$, where the well-known recursion for matrix $\mathbf{P}_{k,n}$ at node k and time instant n is given by [21, 22]:

$$\mathbf{P}_{k,n} = \lambda^{-1} \left(\mathbf{P}_{k,n-1} - \frac{\mathbf{P}_{k,n-1} \mathbf{x}_{k,n} \mathbf{x}_{k,n}^\top \mathbf{P}_{k,n-1}}{\lambda + \mathbf{x}_{k,n}^\top \mathbf{P}_{k,n-1} \mathbf{x}_{k,n}} \right) \tag{10}$$

with the initial condition $\mathbf{P}_{k,0} = \delta^{-1} \mathbf{I}_L$. Specifically, b can be approximated by estimation error at node k and time instant n, i.e.,

$$b = d_{k,n} - \mathbf{x}_{k,n}^\top \mathbf{w}_{k,n-\frac{1}{2}} \approx d_{k,n} - \mathbf{x}_{k,n}^\top \mathbf{w}_{k,n-1} = e_{k,n}. \tag{11}$$

Substituting (2) and (11) into (9), we arrive at the adaptive update step of clustered MT-DRLS algorithm:

$$\begin{aligned} \boldsymbol{\psi}_{k,n} = \mathbf{w}_{k,n-1} &+ \mathbf{P}_{k,n} \mathbf{x}_{k,n} e_{k,n} \\ &- \gamma \sum_{\ell \in \mathcal{N}_k \setminus \mathcal{C}(k)^-} \frac{\rho_{k\ell} + \rho_{\ell k}}{2} (\mathbf{w}_{k,n-1} - \mathbf{w}_{\ell,n-1}). \end{aligned} \tag{12}$$

Let us introduce the entire intermediate estimated vector and the block column matrix with individual entries the $L \times L$ identity matrix, which are defined as follows:

$$\boldsymbol{\psi}_n = \mathrm{col}\{\boldsymbol{\psi}_{1,n}, \ldots, \boldsymbol{\psi}_{K,n}\} \in \mathbb{R}^{KL}, \tag{13}$$

$$\mathbf{H} = \mathrm{col}\{\mathbf{I}_L, \ldots, \mathbf{I}_L\} \in \mathbb{R}^{KL \times L}. \tag{14}$$

We consider the final weighted least-squares problem that requires that each node k communicates with its immediate neighbors within the same cluster [13, 20]:

$$\mathbf{w}_{k,n} = \underset{\mathbf{w} \in \mathbb{R}^L}{\arg\min} \left\{ \left\| \boldsymbol{\psi}_n - \mathbf{H}\mathbf{w} \right\|_{\boldsymbol{\Pi}_k}^2 \right\} \tag{15}$$

with the node-dependent weighting block diagonal matrix $\boldsymbol{\Pi}_k = \mathrm{bdiag}\{a_{1k}\mathbf{I}_L, \ldots, a_{Kk}\mathbf{I}_L\}$. Here, the non-negative combination coefficients $\{a_{\ell k}\}$ are chosen to satisfy [5, 6]:

$$\sum_{\ell \in \mathcal{N}_k \cap \mathcal{C}(k)} a_{\ell k} = 1, \text{ and } \begin{cases} a_{\ell k} > 0, & \text{if } \ell \in \mathcal{N}_k \cap \mathcal{C}(k), \\ a_{\ell k} = 0, & \text{otherwise.} \end{cases} \tag{16}$$

This means that matrix $\mathbf{A}$ with (ℓ, k)-th entry $a_{\ell k}$ is a left-stochastic matrix, i.e., $\mathbf{A}^\top \mathbf{1}_K = \mathbf{1}_K$. Likewise, setting the derivative of (15) with respect to $\mathbf{w}$ to zero, the combination step is given by:

$$\mathbf{w}_{k,n} = \sum_{\ell \in \mathcal{N}_k \cap \mathcal{C}k} a_{\ell k} \boldsymbol{\psi}_{\ell,n}. \tag{17}$$

Therefore, the clustered MT-DRLS algorithm with ATC diffusion strategy is consisting of (12) and (17).

3. TRANSIENT PERFORMANCE ANALYSIS

We now perform the transient performance analysis of MT-DRLS algorithm for clustered multitask networks in the mean and mean-square error sense. The weight error vectors for node k at instant n are defined respectively as follows:

$$\widetilde{\boldsymbol{\psi}}_{k,n} \triangleq \boldsymbol{\psi}_{k,n} - \mathbf{w}_k^\star, \quad \widetilde{\mathbf{w}}_{k,n} \triangleq \mathbf{w}_{k,n} - \mathbf{w}_k^\star. \tag{18}$$

Let $\widetilde{\mathbf{w}}_n$ and $\mathbf{w}^\star$ denote the block weight error vector and the block optimal weight vector, all of size $K \times 1$ with blocks of size $L \times 1$, i.e.,

$$\widetilde{\mathbf{w}}_n \triangleq \mathrm{col}\{\widetilde{\mathbf{w}}_{1,n}, \ldots, \widetilde{\mathbf{w}}_{K,n}\}, \tag{19}$$

$$\mathbf{w}^\star \triangleq \mathrm{col}\{\mathbf{w}_1^\star, \ldots, \mathbf{w}_K^\star\}. \tag{20}$$

We also introduce the following required $K \times K$ block diagonal matrices with each block of size $L \times L$ defined as:

$$\mathbf{R}_{x,n} \triangleq \mathrm{bdiag}\{\mathbf{x}_{1,n}\mathbf{x}_{1,n}^\top, \ldots, \mathbf{x}_{K,n}\mathbf{x}_{K,n}^\top\}, \tag{21}$$

$$\boldsymbol{\Phi}_n \triangleq \mathrm{bdiag}\{\boldsymbol{\Phi}_{1,n}, \ldots, \boldsymbol{\Phi}_{K,n}\}, \tag{22}$$

$$\mathbf{P}_n \triangleq \mathrm{bdiag}\{\mathbf{P}_{1,n}, \ldots, \mathbf{P}_{K,n}\}, \tag{23}$$

$$\boldsymbol{\mathcal{A}} \triangleq \mathbf{A}^\top \otimes \mathbf{I}_L, \tag{24}$$

and the block column vector with individual entries of size $L \times 1$ defined as:

$$\mathbf{s}_{xz,n} \triangleq \mathrm{col}\big\{z_{1,n}\mathbf{x}_{1,n}, \ldots, z_{K,n}\mathbf{x}_{K,n}\big\}. \tag{25}$$

Moreover, it holds that:

$$\mathbb{E}\{\mathbf{s}_{xz,n}\} = \mathbf{0}_{KL} \tag{26}$$

due to the statistical properties of measurement noise $z_{k,n}$. Before proceeding, we introduce the following independence assumption.

Assumption 1. (Independent Regressors): The regression vectors $\mathbf{x}_{k,n}$ arise from a stationary random process that is temporally stationary, temporally white, and spatially independent with the positive-definite covariance matrix $\mathbf{R}_{x,k}$.

A consequence of Assumption 1 is that $\mathbf{x}_{k,n}$ is independent of $\widetilde{\mathbf{w}}_{\ell,m}$ for all ℓ and $m \leq n$. Although not true in general, this assumption is widely used in the theoretical analysis of adaptive filters because it allows to simplify the derivations without constraining the conclusions [21, 22].

3.1. Mean Error Behavior Analysis

With (21) and (22), (6) can be written in the extended form as:

$$\boldsymbol{\Phi}_n = \lambda\boldsymbol{\Phi}_{n-1} + \mathbf{R}_{x,n}. \tag{27}$$

Taking the expectation of both sides, we obtain:

$$\mathbb{E}\{\boldsymbol{\Phi}_n\} = \lambda\mathbb{E}\{\boldsymbol{\Phi}_{n-1}\} + \mathbf{R}_x \tag{28}$$

where the expectation of input correlation matrix $\mathbf{R}_{x,n}$ is given by:

$$\mathbf{R}_x = \mathbb{E}\big\{\mathbf{R}_{x,n}\big\} = \mathrm{bdiag}\big\{\mathbf{R}_{x,1}, \ldots, \mathbf{R}_{x,K}\big\} \in \mathbb{R}^{KL\times KL}. \tag{29}$$

Since matrix $\boldsymbol{\Phi}_n$ (or $\mathbf{P}_n$) only depends on $\mathbf{R}_x$, the relation (28) is very useful in the sequel. In view of (1) and (18), the a priori estimation error given in (11) can be rewritten as:

$$e_{k,n} = z_{k,n} - \mathbf{x}_{k,n}^{\top}\widetilde{\mathbf{w}}_{k,n-1}. \tag{30}$$

Subtracting $\mathbf{w}_k^{\star}$ from both sides of (12) and (17), respectively, then using (18) and (30), we find that:

$$\begin{aligned}\widetilde{\boldsymbol{\psi}}_{k,n} &= \widetilde{\mathbf{w}}_{k,n-1} - \mathbf{P}_{k,n}\mathbf{x}_{k,n}\mathbf{x}_{k,n}^{\top}\widetilde{\mathbf{w}}_{n-1} + \mathbf{P}_{k,n}\mathbf{x}_{k,n}z_{k,n} \\ &\quad + \gamma\sum_{\ell\in\mathcal{N}_k\setminus\mathcal{C}(k)}\frac{\rho_{k\ell}+\rho_{\ell k}}{2}(\mathbf{w}_{\ell,n-1} - \mathbf{w}_{k,n-1}),\end{aligned} \tag{31}$$

$$\widetilde{\mathbf{w}}_{k,n} = \sum_{\ell\in\mathcal{N}_k\cap\mathcal{C}k} a_{\ell k}\widetilde{\boldsymbol{\psi}}_{\ell,n}. \tag{32}$$

Substituting (31) into (32), and using the above definitions (19)–(21) and (23)–(25), the update equation of block weight error vector can be expressed as follows:

$$\widetilde{\mathbf{w}}_n = \boldsymbol{\mathcal{A}}\big[\widetilde{\mathbf{w}}_{n-1} - \mathbf{P}_n\mathbf{R}_{x,n}\widetilde{\mathbf{w}}_{n-1} + \mathbf{P}_n\mathbf{s}_{xz,n} - \gamma\mathbf{Q}(\mathbf{w}_{n-1} + \mathbf{w}^{\star})\big] \tag{33}$$

where

$$\mathbf{Q} = \frac{1}{2}\Big[\mathrm{diag}\Big\{\big(\boldsymbol{\Theta} + \boldsymbol{\Theta}^{\top}\big)\mathbf{1}_K\Big\} - \big(\boldsymbol{\Theta} + \boldsymbol{\Theta}^{\top}\big)\Big] \otimes \mathbf{I}_L. \tag{34}$$

Pre-multiplying both sides of (33) by $\mathbf{P}_n^{-1}\boldsymbol{\mathcal{A}}^{-1}$, using (27) and the relation $\boldsymbol{\Phi}_n = \mathbf{P}_n^{-1}$ based on the definition $\boldsymbol{\Phi}_{k,n} = \mathbf{P}_{k,n}^{-1}$, yields:

$$\boldsymbol{\Phi}_n\boldsymbol{\mathcal{A}}^{-1}\widetilde{\mathbf{w}}_n = \big(\lambda\boldsymbol{\Phi}_{n-1} - \gamma\boldsymbol{\Phi}_n\mathbf{Q}\big)\widetilde{\mathbf{w}}_{n-1} - \gamma\boldsymbol{\Phi}_n\mathbf{Q}\,\mathbf{w}^{\star} + \mathbf{s}_{xz,n}. \tag{35}$$

The aim of the above manipulations is to separate $\mathbf{P}_n$ and $\mathbf{R}_{x,n}$ in the second term on the r.h.s. of (33). Taking the expectation of both sides of (35), and utilizing the property (26), we then obtain:

$$\begin{aligned}\mathbb{E}\big\{\boldsymbol{\Phi}_n\boldsymbol{\mathcal{A}}^{-1}\widetilde{\mathbf{w}}_n\big\} &= \mathbb{E}\big\{\big(\lambda\boldsymbol{\Phi}_{n-1} - \gamma\boldsymbol{\Phi}_n\mathbf{Q}\big)\widetilde{\mathbf{w}}_{n-1}\big\} \\ &\quad - \gamma\mathbb{E}\{\boldsymbol{\Phi}_n\}\mathbf{Q}\,\mathbf{w}^{\star}.\end{aligned} \tag{36}$$

In order to make the analysis mathematical tractable, we need the following approximations [20, 24]:

$$\mathbb{E}\big\{\boldsymbol{\Phi}_n\boldsymbol{\mathcal{A}}^{-1}\widetilde{\mathbf{w}}_n\big\} \approx \mathbb{E}\{\boldsymbol{\Phi}_n\}\boldsymbol{\mathcal{A}}^{-1}\mathbb{E}\{\widetilde{\mathbf{w}}_n\}, \tag{37}$$

$$\mathbb{E}\big\{\boldsymbol{\Phi}_n\widetilde{\mathbf{w}}_n\big\} \approx \mathbb{E}\{\boldsymbol{\Phi}_n\}\mathbb{E}\{\widetilde{\mathbf{w}}_n\}, \tag{38}$$

$$\mathbb{E}\big\{\boldsymbol{\Phi}_n\mathbf{Q}\,\widetilde{\mathbf{w}}_{n-1}\big\} \approx \mathbb{E}\{\boldsymbol{\Phi}_n\}\mathbf{Q}\,\mathbb{E}\{\widetilde{\mathbf{w}}_{n-1}\}. \tag{39}$$

The proofs of (37)–(39) are not presented explicitly due to the limited space, but the simulation results are able to validate their effectiveness and rationality later. Substituting the approximations (37)–(39) into (36), it follows that:

$$\begin{aligned}\mathbb{E}\{\boldsymbol{\Phi}_n\}\boldsymbol{\mathcal{A}}^{-1}\mathbb{E}\{\widetilde{\mathbf{w}}_n\} &= \big(\lambda\mathbb{E}\{\boldsymbol{\Phi}_{n-1}\} - \gamma\mathbb{E}\{\boldsymbol{\Phi}_n\}\mathbf{Q}\big)\mathbb{E}\{\widetilde{\mathbf{w}}_{n-1}\} \\ &\quad - \gamma\mathbb{E}\{\boldsymbol{\Phi}_n\}\mathbf{Q}\,\mathbf{w}^{\star}.\end{aligned} \tag{40}$$

Pre-multiplying both sides of (40) by $\boldsymbol{\mathcal{A}}\mathbb{E}\{\boldsymbol{\Phi}_n\}^{-1}$, it results that

$$\begin{aligned}\mathbb{E}\{\widetilde{\mathbf{w}}_n\} &= \boldsymbol{\mathcal{A}}\big(\lambda\mathbb{E}\{\boldsymbol{\Phi}_n\}^{-1}\mathbb{E}\{\boldsymbol{\Phi}_{n-1}\} - \gamma\mathbf{Q}\big)\mathbb{E}\{\widetilde{\mathbf{w}}_{n-1}\} \\ &\quad - \gamma\boldsymbol{\mathcal{A}}\,\mathbf{Q}\,\mathbf{w}^{\star}\end{aligned} \tag{41}$$

where relation (28) has been used.

3.2. Mean-Square Error Behavior Analysis

The network transient mean-square deviation (MSD) at time instant n is defined by [1, 2]:

$$\mathrm{MSD}_n = \mathrm{tr}\big\{\widetilde{\mathbf{W}}_n\big\}/K \tag{42}$$

with the correlation matrix of block weight error vector over network $\widetilde{\mathbf{W}}_n = \mathbb{E}\big\{\widetilde{\mathbf{w}}_n\widetilde{\mathbf{w}}_n^{\top}\big\}$. In order to investigate the mean-square error behavior of MT-DRLS algorithm, our next aim is to determine the update equation of $\widetilde{\mathbf{W}}_n$. Post-multiplying (35) by its transpose, and taking the expectation of both sides, leads to:

$$\begin{aligned}\mathbf{T}_0 &= \lambda^2\mathbf{T}_1 + \gamma^2\big(\mathbf{T}_2 + \mathbf{T}_3\big) - \gamma\lambda\big(\mathbf{T}_4 + \mathbf{T}_4^{\top}\big) - \gamma\lambda\big(\mathbf{T}_5 + \mathbf{T}_5^{\top}\big) \\ &\quad + \gamma^2\big(\mathbf{T}_6 + \mathbf{T}_6^{\top}\big) + \mathbf{S}_{xz}\end{aligned} \tag{43}$$

where

$$\mathbf{T}_0 = \mathbb{E}\big\{\boldsymbol{\Phi}_n\boldsymbol{\mathcal{A}}^{-1}\widetilde{\mathbf{w}}_n\widetilde{\mathbf{w}}_n^{\top}(\boldsymbol{\mathcal{A}}^{-1})^{\top}\boldsymbol{\Phi}_n\big\}, \tag{44}$$

$$\mathbf{T}_1 = \mathbb{E}\big\{\boldsymbol{\Phi}_{n-1}\widetilde{\mathbf{w}}_{n-1}\widetilde{\mathbf{w}}_{n-1}^{\top}\boldsymbol{\Phi}_{n-1}\big\}, \tag{45}$$

$$\mathbf{T}_2 = \mathbb{E}\big\{\boldsymbol{\Phi}_n\mathbf{Q}\,\widetilde{\mathbf{w}}_{n-1}\widetilde{\mathbf{w}}_{n-1}^{\top}\mathbf{Q}^{\top}\boldsymbol{\Phi}_n\big\}, \tag{46}$$

$$\mathbf{T}_3 = \mathbb{E}\big\{\boldsymbol{\Phi}_n\mathbf{Q}\,\mathbf{w}^{\star}(\mathbf{w}^{\star})^{\top}\mathbf{Q}^{\top}\boldsymbol{\Phi}_n\big\}, \tag{47}$$

$$\mathbf{T}_4 = \mathbb{E}\big\{\boldsymbol{\Phi}_{n-1}\widetilde{\mathbf{w}}_{n-1}\widetilde{\mathbf{w}}_{n-1}^{\top}\mathbf{Q}^{\top}\boldsymbol{\Phi}_n\big\}, \tag{48}$$

$$\mathbf{T}_5 = \mathbb{E}\big\{\boldsymbol{\Phi}_{n-1}\widetilde{\mathbf{w}}_{n-1}(\mathbf{w}^{\star})^{\top}\mathbf{Q}^{\top}\boldsymbol{\Phi}_n\big\}, \tag{49}$$

$$\mathbf{T}_6 = \mathbb{E}\big\{\boldsymbol{\Phi}_n\mathbf{Q}\,\widetilde{\mathbf{w}}_{n-1}(\mathbf{w}^{\star})^{\top}\mathbf{Q}^{\top}\boldsymbol{\Phi}_n\big\}, \tag{50}$$

$$\mathbf{S}_{xz} = \mathbb{E}\big\{\mathbf{s}_{xz,n}\mathbf{s}_{xz,n}^{\top}\big\}. \tag{51}$$

For mathematical tractability of analysis, we introduce the following necessary approximations:

$$\begin{aligned}\mathbf{T}_0 &= \mathbb{E}\big\{\boldsymbol{\Phi}_n\boldsymbol{\mathcal{A}}^{-1}\widetilde{\mathbf{w}}_n\widetilde{\mathbf{w}}_n^{\top}(\boldsymbol{\mathcal{A}}^{-1})^{\top}\boldsymbol{\Phi}_n\big\} \\ &\approx \mathbb{E}\{\boldsymbol{\Phi}_n\}\boldsymbol{\mathcal{A}}^{-1}\widetilde{\mathbf{W}}_n(\boldsymbol{\mathcal{A}}^{-1})^{\top}\mathbb{E}\{\boldsymbol{\Phi}_n\},\end{aligned} \tag{52}$$

$$\mathbf{T}_1 = \mathbb{E}\{\boldsymbol{\Phi}_{n-1}\widetilde{\mathbf{w}}_{n-1}\widetilde{\mathbf{w}}_{n-1}^{\top}\boldsymbol{\Phi}_{n-1}\} \tag{53}$$
$$\approx \mathbb{E}\{\boldsymbol{\Phi}_{n-1}\}\widetilde{\mathbf{W}}_{n-1}\mathbb{E}\{\boldsymbol{\Phi}_{n-1}\},$$
$$\mathbf{T}_2 = \mathbb{E}\{\boldsymbol{\Phi}_n\mathbf{Q}\,\widetilde{\mathbf{w}}_{n-1}\widetilde{\mathbf{w}}_{n-1}^{\top}\mathbf{Q}^{\top}\boldsymbol{\Phi}_n\} \tag{54}$$
$$\approx \mathbb{E}\{\boldsymbol{\Phi}_n\}\mathbf{Q}\,\widetilde{\mathbf{W}}_{n-1}\mathbf{Q}^{\top}\mathbb{E}\{\boldsymbol{\Phi}_n\},$$
$$\mathbf{T}_3 = \mathbb{E}\{\boldsymbol{\Phi}_n\mathbf{Q}\,\mathbf{w}^{\star}(\mathbf{w}^{\star})^{\top}\mathbf{Q}^{\top}\boldsymbol{\Phi}_n\} \tag{55}$$
$$\approx \mathbb{E}\{\boldsymbol{\Phi}_n\}\mathbf{Q}\,\mathbf{w}^{\star}(\mathbf{w}^{\star})^{\top}\mathbf{Q}^{\top}\mathbb{E}\{\boldsymbol{\Phi}_n\},$$
$$\mathbf{T}_4 = \mathbb{E}\{\boldsymbol{\Phi}_{n-1}\widetilde{\mathbf{w}}_{n-1}\widetilde{\mathbf{w}}_{n-1}^{\top}\mathbf{Q}^{\top}\boldsymbol{\Phi}_n\} \tag{56}$$
$$\approx \mathbb{E}\{\boldsymbol{\Phi}_{n-1}\}\widetilde{\mathbf{W}}_{n-1}\mathbf{Q}^{\top}\mathbb{E}\{\boldsymbol{\Phi}_n\},$$
$$\mathbf{T}_5 = \mathbb{E}\{\boldsymbol{\Phi}_{n-1}\widetilde{\mathbf{w}}_{n-1}(\mathbf{w}^{\star})^{\top}\mathbf{Q}^{\top}\boldsymbol{\Phi}_n\} \tag{57}$$
$$\approx \mathbb{E}\{\boldsymbol{\Phi}_{n-1}\}\mathbb{E}\{\widetilde{\mathbf{w}}_{n-1}\}(\mathbf{w}^{\star})^{\top}\mathbf{Q}^{\top}\mathbb{E}\{\boldsymbol{\Phi}_n\},$$
$$\mathbf{T}_6 = \mathbb{E}\{\boldsymbol{\Phi}_n\mathbf{Q}\,\widetilde{\mathbf{w}}_{n-1}(\mathbf{w}^{\star})^{\top}\mathbf{Q}^{\top}\boldsymbol{\Phi}_n\} \tag{58}$$
$$\approx \mathbb{E}\{\boldsymbol{\Phi}_n\}\mathbf{Q}\,\mathbb{E}\{\widetilde{\mathbf{w}}_{n-1}\}(\mathbf{w}^{\star})^{\top}\mathbf{Q}^{\top}\mathbb{E}\{\boldsymbol{\Phi}_n\}.$$

The corresponding proofs of (52)–(58) are omitted due to space constraints. The effectiveness and rationality can be testified by the simulation results in the next section. According to assumption 1 and the statistical property of measurement noise $z_{k,n}$, the matrix $\mathbf{S}_{xz}$ can be determined as follows:

$$\mathbf{S}_{xz} = \text{bdiag}\{\sigma_{z,1}^2\mathbf{R}_{x,1},\ldots,\sigma_{z,K}^2\mathbf{R}_{x,K}\} = \boldsymbol{\Sigma}_z\mathbf{R}_x \tag{59}$$

with block diagonal matrix $\boldsymbol{\Sigma}_z = \text{bdiag}\{\sigma_{z,1}^2\mathbf{I}_L,\ldots,\sigma_{z,K}^2\mathbf{I}_L\}$. Substituting (52)–(59) into (43), then multiplying from the left by $\boldsymbol{\mathcal{A}}\mathbb{E}\{\boldsymbol{\Phi}_n\}^{-1}$ and multiplying from the right by $\mathbb{E}\{\boldsymbol{\Phi}_n\}^{-1}\boldsymbol{\mathcal{A}}^{\top}$ simultaneously, we finally arrive at the recursion of $\widetilde{\mathbf{W}}_n$ as follows:

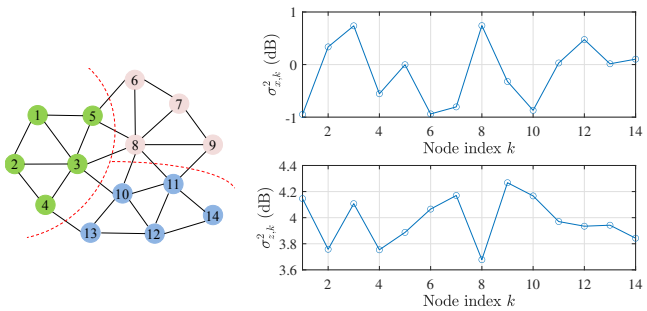
$$\begin{aligned}\widetilde{\mathbf{W}}_n = \boldsymbol{\mathcal{A}}\Big[&\lambda^2\mathbb{E}\{\boldsymbol{\Phi}_n\}^{-1}\mathbb{E}\{\boldsymbol{\Phi}_{n-1}\}\widetilde{\mathbf{W}}_{n-1}\mathbb{E}\{\boldsymbol{\Phi}_{n-1}\}\mathbb{E}\{\boldsymbol{\Phi}_n\}^{-1}\\
&+\gamma^2\big(\mathbf{Q}\,\widetilde{\mathbf{W}}_{n-1}\mathbf{Q}^{\top}+\mathbf{Q}\,\mathbf{w}^{\star}(\mathbf{w}^{\star})^{\top}\mathbf{Q}^{\top}\big)\\
&-\gamma\lambda\big(\mathbb{E}\{\boldsymbol{\Phi}_n\}^{-1}\mathbb{E}\{\boldsymbol{\Phi}_{n-1}\}\widetilde{\mathbf{W}}_{n-1}\mathbf{Q}^{\top}\\
&+\mathbf{Q}\,\widetilde{\mathbf{W}}_{n-1}^{\top}\mathbb{E}\{\boldsymbol{\Phi}_{n-1}\}\mathbb{E}\{\boldsymbol{\Phi}_n\}^{-1}\big)\\
&-\gamma\lambda\big(\mathbb{E}\{\boldsymbol{\Phi}_n\}^{-1}\mathbb{E}\{\boldsymbol{\Phi}_{n-1}\}\mathbb{E}\{\widetilde{\mathbf{w}}_{n-1}\}(\mathbf{w}^{\star})^{\top}\mathbf{Q}^{\top}\\
&+\mathbf{Q}\,\mathbf{w}^{\star}\mathbb{E}\{\widetilde{\mathbf{w}}_{n-1}\}^{\top}\mathbb{E}\{\boldsymbol{\Phi}_{n-1}\}\mathbb{E}\{\boldsymbol{\Phi}_n\}^{-1}\big)\\
&+\gamma^2\big(\mathbf{Q}\,\mathbb{E}\{\widetilde{\mathbf{w}}_{n-1}\}(\mathbf{w}^{\star})^{\top}\mathbf{Q}^{\top}+\mathbf{Q}\,\mathbf{w}^{\star}\mathbb{E}\{\widetilde{\mathbf{w}}_{n-1}\}^{\top}\mathbf{Q}^{\top}\big)\\
&+\mathbb{E}\{\boldsymbol{\Phi}_n\}^{-1}\boldsymbol{\Sigma}_z\mathbf{R}_x\mathbb{E}\{\boldsymbol{\Phi}_n\}^{-1}\Big]\boldsymbol{\mathcal{A}}^{\top}.\end{aligned} \tag{60}$$

It should be pointed out that the above recursive evaluation needs to employ relation (28) and (41). By (60), we can characterize the transient mean-square errors of the clustered MT-DRLS algorithm.

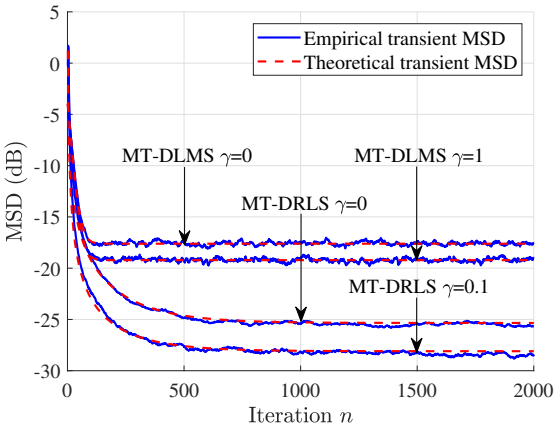
4. NUMERICAL TESTS

In this section, we provide an illustrative example to show the superior performance of clustered MT-DRLS algorithm, and to validate the obtained transient analytical models. We considered a connected network consisting of 14 nodes grouped into 3 clusters shown in Fig. 1(a). The optimal weight vectors to be estimated in each cluster were $\mathbf{w}_{\mathcal{C}_1}^{\star} = [0.5196, -0.3667]^{\top}$, $\mathbf{w}_{\mathcal{C}_2}^{\star} = [0.4952, -0.3783]^{\top}$, and $\mathbf{w}_{\mathcal{C}_3}^{\star} = [0.4951, -0.4079]^{\top}$, respectively. The regression vectors $\mathbf{x}_{k,n}$ were zero-mean random vectors governed by a Gaussian distribution with covariance matrix $\mathbf{R}_{x,k} = \sigma_{x,k}\mathbf{I}_L$. The measurement noise $z_{k,n}$ was i.i.d. Gaussian with zero-mean and variances $\sigma_{z,k}^2$. The variances $\sigma_{x,k}^2$ and $\sigma_{z,k}^2$ are depicted in Fig. 1 (b), respectively. Each combination coefficient $a_{k\ell}$ was chosen as $|\mathcal{N}_k \cap \mathcal{C}(k)|^{-1}$ for all $\ell \in \mathcal{N}_k \cap \mathcal{C}(k)$, where $|\cdot|$ denotes the cardinality of its argument. The regularization weight $\rho_{k\ell}$ was uniformly chosen as $\rho_{k\ell} = |\mathcal{N}_k \backslash \mathcal{C}(k)|^{-1}$ for $\ell \in \mathcal{N}_k \backslash \mathcal{C}(k)$, and $\rho_{k\ell} = 0$ for any other ℓ. The step-sizes of MT-DLMS algorithm were all set to 0.03. The forgetting factor λ and the initialization parameter δ of clustered MT-DRLS algorithm were set to 0.995 and 0.05, respectively. All the empirical learning curves were obtained by averaging over 200 Monte-Carlo runs.

Fig. 1(c) shows that the clustered MT-DRLS algorithms significantly outperforms the counterpart clustered MT-DLMS algorithms in terms of convergence rate, steady-state errors, and parameter estimation accuracy. As shown in Fig. 1(c), the MT-DRLS algorithm with parameter $\gamma = 0.1$ gains about 3 dB over the MT-DRLS algorithm with parameter $\gamma = 0$ in the steady-state MSD, and needs about 300 iterations less before attaining the steady-state phase of MT-DRLS algorithm with parameter $\gamma = 0$. More importantly, we can also see that the consistent agreement between empirical and theoretical MSD curves validates the accuracy and effectiveness of the transient theoretical analysis. Last, this consistency also validates all the necessary approximations introduced in the analysis.



(a) Network topology. (b) Input and noise variances.



(c) Empirical vs. theoretical network MSD.

Fig. 1. Network setup and simulation results.

5. CONCLUSION

In this paper, we presented the DRLS algorithm with ATC diffusion strategy over clustered multitask networks to improve the performance of clustered MT-DLMS algorithm. We also provided a transient analysis of the algorithm in the mean and mean-square error sense. In future works, we will study its steady-state behavior.

6. REFERENCES

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

[2] A. H. Sayed, "Diffusion adaptation over networks," in *Academic Press Library in Signal Processing*, vol. 3, pp. 323–453. Elsevier, 2014.

[3] P. M. Djuric and C. Richard, *Cooperative and Graph Signal Processing: Principles and Applications*, Academic Press, New York, USA, 2018.

[4] J. Chen, C. Richard, and A. H. Sayed, "Diffusion LMS for clustered multitask networks," in *Proc. IEEE ICASSP*, May 2014, pp. 5487–5491.

[5] 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.

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

[7] J. Chen, C. Richard, and A. H. Sayed, "Multitask diffusion adaptation over networks with common latent representations," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 3, pp. 563–579, Apr. 2017.

[8] R. Nassif, C. Richard, A. Ferrari, and A. H. Sayed, "Diffusion LMS for multitask problems with local linear equality constraints," *IEEE Trans. Signal Process.*, vol. 65, no. 19, pp. 4979–4993, Oct. 2017.

[9] R. Nassif, C. Richard, A. Ferrari, and A. H. Sayed, "Proximal multitask learning over networks with sparsity-inducing coregularization," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6329–6344, Dec. 2016.

[10] D. Jin, J. Chen, C. Richard, and J. Chen, "Online proximal learning over jointly sparse multitask networks with $\ell_{\infty,1}$ regularization," *IEEE Trans. on Signal Process.*, vol. 68, pp. 6319–6335, 2020.

[11] V. C. Gogineni, S. P. Talebi, and S. Werner, "Performance of clustered multitask diffusion LMS suffering from inter-node communication delays," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 68, no. 7, pp. 2695–2699, Jul. 2021.

[12] R. Nassif, C. Richard, J. Chen, and A. H. Sayed, "Multitask learning over graphs: An approach for distributed, streaming machine learning," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 14–25, May 2020.

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

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

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

[16] Z. Liu, Y. Liu, and C. Li, "Distributed sparse recursive least-squares over networks," *IEEE Trans. on Signal Process.*, vol. 62, no. 6, pp. 1386–1395, Mar. 2014.

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

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

[19] A. Rastegarnia, "Reduced-communication diffusion RLS for distributed estimation over multi-agent networks," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 67, no. 1, pp. 177–181, Jan. 2020.

[20] W. Gao, J. Chen, and C. Richard, "Transient theoretical analysis of diffusion RLS algorithm for cyclostationary colored inputs," *IEEE Signal Process. Lett.*, vol. 28, pp. 1160–1164, 2021.

[21] S. Haykin, *Adaptive Filter Theory*, Prentice-Hall, New Jersey, 2nd edition, 1991.

[22] A. H. Sayed, *Fundamentals of Adaptive Filtering*, Wiley, New York, 2003.

[23] X. Cao and K. J. R. Liu, "Decentralized sparse multitask RLS over networks," *IEEE Trans. on Signal Process.*, vol. 65, no. 23, pp. 6217–6232, Dec. 2017.

[24] E. Eweda, N. J. Bershad, and J. C. M. Bermudez, "Stochastic analysis of the recursive least squares algorithm for cyclostationary colored inputs," *IEEE Trans. on Signal Process.*, vol. 68, pp. 676–686, 2020.