Multitask Diffusion Adaptation Over Networks With Common Latent Representations

Jie Chen, Cédric Richard, Ali Sayed

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2017.2671789, IEEE Journal of Selected Topics in Signal Processing

Multitask diffusion adaptation over networks with common latent representations

Jie Chen, *Member, IEEE*, Cédric Richard, *Senior Member, IEEE*, Ali H. Sayed, *Fellow Member, IEEE*

***Abstract*—Online learning with streaming data in a distributed and collaborative manner can be useful in a wide range of applications. This topic has been receiving considerable attention in recent years with emphasis on both single-task and multitask scenarios. In single-task adaptation, agents cooperate to track an objective of common interest, while in multitask adaptation agents track multiple objectives simultaneously. Regularization is one useful technique to promote and exploit similarity among tasks in the latter scenario. This work examines an alternative way to model relations among tasks by assuming that they all share a common latent feature representation. As a result, a new multitask learning formulation is presented and algorithms are developed for its solution in a distributed online manner. We present a unified framework to analyze the mean-square-error performance of the adaptive strategies, and conduct simulations to illustrate the theoretical findings and potential applications.**

***Index Terms*—Multitask learning, distributed optimization, common latent subspace, online adaptation, diffusion strategy, collaborative processing, performance analysis.**

I. Introduction

Multi-agent networks usually consist of a large number of interconnected agents or nodes. Interconnections between the agents allow them to share information and collaborate in order to solve complex tasks collectively. Examples abound in the realm of social, economic and biological networks. Distributed algorithms over such networks offer a valuable alternative to centralized solutions with useful properties such as scalability, robustness, and decentralization. When endowed with adaptation abilities, these algorithms enable agents to continuously learn and adapt in an online manner to concept drifts in their data streams [2], [3]. Broadly, distributed strategies for online parameter estimation can be applied to single-task or multi-task scenarios. In the first case, agents cooperate with each other to estimate a single parameter vector of interest, such as tracking a common target. Reaching consensus among the agents is critical for successful inference in these problems. In the multitask case, the agents cooperate to estimate multiple parameter vectors simultaneously, such as tracking a collection of targets moving in formation [4].

The work of J. Chen was supported in part by the NSFC grant 61671382. The work of C. Richard was supported in part by the Agence Nationale pour la Recherche, France, (ODISSEE project, ANR-13-ASTR-0030). The work of A. H. Sayed was supported in part by NSF grants CIF-1524250 and ECCS-1407712. A short and preliminary version of this work appears in the conference publication [1]. J. Chen is with CIAIC of School of Marine Science and Technology, Northwestern Polytechinical University, Xi'an, 710072, China (e-mail: dr.jie.chen@ieee.org). C. Richard is with the Université de Nice Sophia-Antipolis, CNRS, France (e-mail: cedric.richard@unice.fr), in collaboration with Morpheme team (INRIA Sophia-Antipolis). A. H. Sayed is with the department of electrical engineering, University of California, Los Angeles, CA 9005-1594, USA (email: sayed@ee.ucla.edu).

Extensive studies have been conducted on adaptive distributed strategies for single-task problems. Existing techniques include incremental [5]–[8], consensus [9]–[11], and diffusion strategies [1], [2], [12]–[17]. Incremental techniques require determining a cyclic path that runs across all nodes, which is generally a challenging (NP-hard) task to perform. Besides, feature makes the incremental strategies sensitive to link failures and problematic for adaptation. Consensus techniques aim to reach an agreement among nodes on the estimate of interest via local information exchanges, but they have been shown [2], [3] to suffer from instability problems when used in the context of adaptive networks due to an inherent asymmetry in the update equations. Diffusion techniques, on the other hand, have been shown to have superior stability and performance ranges [18] than consensus-based implementations. For these reasons, we shall focus on diffusion-type implementations in this paper.

Besides single-task scenarios, there are also applications where it is desirable to estimate multiple parameter vectors at the same time, rather than promote consensus among all agents [19]. For example, geosensor networks that monitor dynamic spatial fields, such as temperature or windspeed variations in geographic environments, require node-specific estimation problems that are able to take advantage of the spatial correlation between the measurements of neighboring nodes [20], [21]. A second example is the problem of collaborative target tracking where agents track several objects simultaneously [4], [19]. Motivated by these applications, there have been several variations of distributed strategies to deal with multitask scenarios as well. Existing strategies mostly depend on how the tasks relate to each other and on exploiting some prior information. In a first scenario, nodes are grouped into clusters, and each cluster of nodes is interested in estimating its own parameter vector. Although clusters may generally have distinct though related estimation tasks to perform, the nodes may still be able to capitalize on inductive transfer between clusters to improve their estimation accuracy. Multitask diffusion strategies were developed to perform estimation under these conditions [4], [22]. One useful way to do so is to employ regularization. A couple of other useful works have also addressed variations of this scenario where the only available information is that clusters may exist in the network but nodes do not know which other nodes share the same estimation task [23]–[25]. In [26], the authors use multitask diffusion adaptation with a node clustering strategy to identify a model between the gait information and electroencephalographic signals. In [27], the authors consider the framework in [4] to devise a distributed strategy that allows each node in the network to locally adapt inter-cluster cooperation weights. The authors in [28] promote cooperation

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2017.2671789, IEEE Journal of Selected Topics in Signal Processing

between clusters with ℓ_1-norm co-regularizers. They derive a closed-form expression of the proximal operator, and introduce a strategy that also allows each node to automatically set its inter-cluster cooperation weights. The works in [29], [30] propose alternative node clustering strategies. In a second scenario, it is assumed that there are parameters of global interest to all nodes in the network, a collection of parameters of common interest within sub-groups of nodes, and a set of parameters of local interest at each node. A diffusion strategy was developed to perform estimation under these conditions [31], [32]. Likewise, in the works [33]–[35], distributed algorithms are derived to estimate node-specific parameter vectors that lie in a common latent signal subspace. In another work [36], the diffusion LMS algorithm is extended to deal with structured criteria built upon groups of variables, leading to a flexible framework that can encode various structures in the parameters. An unsupervised strategy to differentially promote or inhibit collaboration between nodes depending on their group is also introduced.

Alternatively, in recent years, there has been an increasing interest in modeling relations between tasks by assuming that all tasks share a common feature representation in a latent subspace [37]–[39]. The authors in [38] proposed a non-convex method based on Alternating Structure Optimization (ASO) for identifying the task structure. A convex relaxation of this approach was developed in [40]. In [39], the authors showed the equivalence between ASO, clustered multitask learning [41], [42] and their convex relaxations. The efficiency of such task relationships has been demonstrated in these works for clustering and classification problems. In our preliminary work [1], we introduced this framework within the context of distributed online adaptation over networks. Useful applications can be envisaged. First, consider the case where the common subspace is spanned by certain selected columns of the identity matrix. This means that a subset of the entries of the parameter vector to be estimated are common to all nodes while no further restriction is imposed on the other entries. Another example concerns beamforming for antenna arrays with a generalized side-lobe canceller (GSC). The latent subspace corresponds to the space where interfering signals reside [43]. A third example deals with cooperative spectrum sensing in cognitive radios, where the common latent subspace characterizes common interferers [31].

Drawing on these motivations, this paper deals with distributed learning and adaptation over multitask networks with common latent representation subspaces. Algorithms are designed accordingly, and their performance analyzed. The contributions of this work include the following main aspects:

- We formulate a new multitask estimation problem, which assumes that all tasks share a common latent subspace representation in addition to node-specific contributions. Additional constraints can be incorporated if needed. This work contrasts with earlier works [4], [28], where the inductive transfer between learning tasks is promoted by regularizers. It also differs from [31], which considers direct models by stacking local and global variables in an augmented parameter vector. Moreover, the work [38] uses a similar inductive transfer model but the common latent subspace is unknown and embedded into a joint estimation process. Our work is the first one to introduce an online estimation algorithm over networks. Estimating the common latent subspace of interest within this context is a challenging perspective.
- We explain how this formulation can be tailored to fit individual application contexts by considering additional model constraints. We illustrate this fact by considering two convex optimization problems and the associated distributed online algorithms. The first algorithm is a generalization in some sense of the diffusion LMS algorithm, which can be retrieved by defining the low-dimensional common latent subspace as the whole parameter space. The second algorithm uses ℓ_2-norm regularization to account for the multitask nature of the problem. This opens the way to other regularization schemes depending on the application at hand.
- We present a unified framework for analyzing the performance of these algorithms. This framework also allows to address the performance analysis of the multitask algorithms in [4], [19], [44], [45] in a generic manner, though these analyses were performed independently of each other in these works.

The rest of the paper is organized as follows. Section II introduces the multitask estimation problem considered in this paper. Then, two distributed learning strategies are derived in Section III by imposing different constraints on common and node-specific representation subspaces. Section IV provides a general framework for analyzing distributed algorithms of this form. In Section V, experiments are conducted to illustrate the characteristics of these algorithms. Section VI concludes the paper and connects our work with several other learning strategies.

Notation. Normal font x denotes scalars. Boldface small letters $\boldsymbol{x}$ denote vectors. All vectors are column vectors. Boldface capital letters $\boldsymbol{X}$ denote matrices. The asterisk $(\cdot)^*$ denotes complex conjugation for scalars and complex-conjugate transposition for matrices. The superscript $(\cdot)^\top$ represents transpose of a matrix or a vector, and $\|\cdot\|$ is the ℓ_2-norm of its matrix or vector argument. $\mathrm{Re}\{\cdot\}$ and $\mathrm{Im}\{\cdot\}$ denote the real and imaginary parts of their complex argument, respectively. Matrix trace is denoted by $\mathrm{trace}(\cdot)$. The operator $\mathrm{col}\{\cdot\}$ stacks its vector arguments on the top of each other to generate a connected vector. The operator $\mathrm{diag}\{\cdot\}$ formulates a (block) diagonal matrix with its arguments. Identity matrix of size $N \times N$ is denoted by $\boldsymbol{I}_N$. Kronecker product is denoted by $\otimes$, and expectation is denoted by $\mathbb{E}\{\cdot\}$. We denote by $\mathcal{N}_k$ the set of node indices in the neighborhood of node k, including k itself, and $|\mathcal{N}_k|$ its set cardinality.

II. Matched subspace estimation over multitask networks

A. Multitask estimation problems over networks

Consider a connected network composed of N nodes. The problem is to estimate an $L \times 1$ unknown vector $\boldsymbol{w}_k^o$ at each node k from collected measurements. At each time n, node k has access to local streaming measurements $\{d_k(n), \boldsymbol{x}_{k,n}\}$,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2017.2671789, IEEE Journal of Selected Topics in Signal Processing

where $d_k(n)$ is a scalar zero-mean reference signal, and $\boldsymbol{x}_{k,n}$ is a $1 \times L$ zero-mean row regression vector with covariance matrix $\boldsymbol{R}_{x,k} = \mathbb{E}\{\boldsymbol{x}_{k,n}^* \boldsymbol{x}_{k,n}\} > 0$. The data at agent k and time n are assumed to be related via the linear model:

$$d_k(n) = \boldsymbol{x}_{k,n}\boldsymbol{w}_k^o + z_k(n) \tag{1}$$

where $\boldsymbol{w}_k^o$ is an unknown complex parameter vector, and $z_k(n)$ is a zero-mean i.i.d. noise with variance $\sigma_{z,k}^2 = \mathbb{E}\{|z_k(n)|^2\}$. The noise signal $z_k(n)$ is assumed to be independent of any other signal. Let $J_k(\boldsymbol{w})$ be a differentiable convex cost function at agent k. In this paper, we shall consider the mean-square-error criterion:

$$J_k(\boldsymbol{w}) = \mathbb{E}\{|d_k(n) - \boldsymbol{x}_{k,n}\boldsymbol{w}|^2\} \tag{2}$$

It is clear from (1) that each $J_k(\boldsymbol{w})$ is minimized at $\boldsymbol{w}_k^o$. We refer to each parameter $\boldsymbol{w}_k^o$ to estimate (or model in a more general sense) as a task. Depending on whether the minima of all $J_k(\boldsymbol{w})$ are achieved at the same $\boldsymbol{w}_k^o$ or not, the distributed learning problem can be single-task or multitask oriented [4].

With single-task networks, all agents aim at estimating the same parameter vector $\boldsymbol{w}^o$ shared by the entire network, that is,

$$\boldsymbol{w}_k^o = \boldsymbol{w}^o \tag{3}$$

for all $k \in \{1, ..., N\}$. Several popular collaborative strategies, such as diffusion LMS [1], [2], [13], [14], were derived to address this problem by seeking the minimizer of the following aggregate cost function:

$$J^{\text{glob}}(\boldsymbol{w}) = \sum_{k=1}^{N} J_k(\boldsymbol{w}) \tag{4}$$

in a distributed manner. Since the individual costs (2) admit the same solution, $\boldsymbol{w}^o$ is also the solution of (4). It has been shown that using proper cooperative strategies to solve (4) can improve the estimation performance [2], [3].

With multitask networks, each agent aims at determining a local parameter vector $\boldsymbol{w}_k^o$. It is assumed that some similarities or relations exist among the parameter vectors of neighboring agents so that cooperation can still be meaningful, namely,

$$\boldsymbol{w}_k^o \sim \boldsymbol{w}_\ell^o \;\text{ if } \ell \in \mathcal{N}_k \tag{5}$$

where the symbol $\sim$ refers to a similarity relationship in some sense, which can be exploited to enhance performance. Depending on the problem characteristics, this property can be promoted in several ways, e.g., by introducing some regularization term, or by assuming a common latent structure. Networks may also be structured into clusters where agents within each cluster estimate the same parameter vector [4], [44].

B. Node-specific subspace constraints

Although agents aim to estimate distinct minimizers $\boldsymbol{w}_k^o$, exploiting relationships between solutions can make cooperation among agents beneficial. Regularization is one popular technique for introducing prior information about the solution. It can improve estimation accuracy though it may introduce bias [4], [19], [46]. In this paper, we explore an alternative strategy that assumes that the hypothesis spaces partially overlap. Specifically, we assume that each $\boldsymbol{w}_k^o$ can be expressed in the form:

$$\boldsymbol{w}_k^o = \boldsymbol{\Theta}\boldsymbol{u}^o + \boldsymbol{\epsilon}_k^o \tag{6}$$

where $\boldsymbol{\Theta}\boldsymbol{u}^o$ is common to all nodes with $\boldsymbol{\Theta}$ denoting an $L \times M$ matrix with known entries and $\boldsymbol{u}^o$ an unknown $M \times 1$ parameter vector (common to all nodes), and where $\boldsymbol{\epsilon}_k^o$ is an unknown node-specific component. We assume that matrix $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M]$ is full-rank with $M \leq L$. Overcomplete sets of column vectors $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$ may be advantageous in some scenarios but this usually requires to impose further constraints such as sparsity over $\boldsymbol{u}^o$. We shall not discuss this case further in order to focus on the main points of the presentation. Model (6) means that all tasks share the same parameter vector $\boldsymbol{\Theta}\boldsymbol{u}^o$, which lies in the subspace spanned by columns of $\boldsymbol{\Theta}$. This subspace representation can be useful in several applications. For instance, consider the case where $\boldsymbol{\Theta}$ is composed of selected columns of the identity matrix $\boldsymbol{I}_L$. This means that a subset of the entries of $\boldsymbol{w}_k^o$ are common to all agents while no further assumptions are imposed on the other entries. This situation is a natural generalization of the single-task scenario. Another example concerns beamforming problems with a generalized sidelobe canceller (GSC), where $\boldsymbol{\Theta}$ acts as a blocking matrix to cancel signal components that lie in the constraint space [43]. In machine learning, formulation (6) is referred to as the alternating structure optimization (ASO) problem [38], [39]. The subspace $\boldsymbol{\Theta}$ is, however, learnt simultaneously via a non-convex optimization procedure. In what follows, we shall assume that $\boldsymbol{\Theta}$ is known by each agent.

Before proceeding further, we clarify the difference between model (6) addressed here and in our preliminary work [1], and the model studied in [31], [32], [35], [47], [48]. In these last works, the authors consider particular information access models where global and local components are assumed to be related to distinct regressors. The centralized problem can then be formulated by stacking the global and local regressors, and by considering a parameter vector augmented accordingly. In our work, motivated by applications of the latent space model in batch-mode learning, we address the problem where the parameter vectors to be estimated lie in global and local latent subspaces. We do not need to distinguish explicitly between global and local regressors. Instead, as shown in the sequel, some extra conditions are needed so that model (6) is identifiable. Among other possibilities, we shall investigate two strategies where constraints on $\boldsymbol{\Theta}$ and $\boldsymbol{\epsilon}_k^o$ are imposed.

Replacing (6) into (2), the global cost function is expressed as a function of a common parameter $\boldsymbol{u}$ and node-specific perturbations $\{\boldsymbol{\epsilon}_k\}_{k=1}^N$:

$$J^{\text{glob}}\left(\boldsymbol{u}, \{\boldsymbol{\epsilon}_k\}_{k=1}^N\right) = \sum_{k=1}^{N} \mathbb{E}\left\{|d_k(n) - \boldsymbol{x}_{k,n}(\boldsymbol{\Theta}\boldsymbol{u} + \boldsymbol{\epsilon}_k)|^2\right\} \tag{7}$$

We expect the estimation of $\boldsymbol{w}_k^o$ by each agent to benefit from the cooperative estimation of $\boldsymbol{u}$. Problem (7) is still insufficient for estimating the tasks $\{\boldsymbol{w}_k^o\}$. This is because the decomposition $\boldsymbol{w}_k = \boldsymbol{\Theta}\,\boldsymbol{u} + \boldsymbol{\epsilon}_k$ is not unique. Indeed, given any optimum solution $\{\bar{\boldsymbol{u}}, \bar{\boldsymbol{\epsilon}}_k\}$, and any $\boldsymbol{s} = \boldsymbol{\Theta}\boldsymbol{x}$, we can

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2017.2671789, IEEE Journal of Selected Topics in Signal Processing

generate another optimum solution by considering the shift $\{\bar{\boldsymbol{u}} - \boldsymbol{x}, \bar{\boldsymbol{\epsilon}}_k + \boldsymbol{s}\}$. This ambiguity prevents us from deriving collaboration strategies based on $\boldsymbol{u}$. From the point of view of convex analysis, the Hessian matrix of (7) is rank deficient and no unique solution exists.

III. Problem formulations and solution algorithms

Problem (7) can be modified to make it well-determined and more meaningful. In this section, among other possibilities, we investigate two strategies that consist of imposing further constraints and derive the corresponding distributed algorithms. These two formulations guarantee the uniqueness of the solution and have clear interpretations.

A. Node-specific subspace constraints

We restrict the node-specific components $\{\boldsymbol{\epsilon}_k\}_{k=1}^N$ to lie in the complementary subspace to span($\boldsymbol{\Theta}$). The problem can be formulated as:

$$\begin{aligned}\min_{\boldsymbol{u},\{\boldsymbol{\epsilon}_k\}_{k=1}^N} \; & J^{\text{glob}}\big(\boldsymbol{u}, \{\boldsymbol{\epsilon}_k\}_{k=1}^N\big) \\ & \text{subject to } \boldsymbol{\epsilon}_k \in \text{span}(\boldsymbol{\Theta}_\perp), \quad \forall k = 1, \ldots, N\end{aligned} \tag{8}$$

where the $L - M$ columns of matrix $\boldsymbol{\Theta}_\perp$ span the complementary subspace to span($\boldsymbol{\Theta}$), that is, $\boldsymbol{\Theta}^* \boldsymbol{\Theta}_\perp = \boldsymbol{0}$. We write:

$$\boldsymbol{\epsilon}_k = \boldsymbol{\Theta}_\perp \, \boldsymbol{\xi}_k \tag{9}$$

where $\boldsymbol{\xi}_k$ is a column vector of size $(L - M)$. Now, replacing (9) into (8), the optimization problem becomes unconstrained and the objective function is given by:

$$\begin{aligned}&J^{\text{glob}}\big(\boldsymbol{u}, \{\boldsymbol{\xi}_k\}_{k=1}^N\big) \\ &= \sum_{k=1}^N \mathbb{E}\big\{|d_k(n) - \boldsymbol{x}_{k,n}(\boldsymbol{\Theta}\boldsymbol{u} + \boldsymbol{\Theta}_\perp \boldsymbol{\xi}_k)|^2\big\} \\ &= \sum_{k=1}^N \mathbb{E}\{|d_k(n)|^2\} + \boldsymbol{u}^*\boldsymbol{\Theta}^*\Big(\sum_{k=1}^N \boldsymbol{R}_{x,k}\Big)\boldsymbol{\Theta}\boldsymbol{u} \\ &+ \sum_{k=1}^N \boldsymbol{\xi}_k^* \boldsymbol{\Theta}_\perp^* \boldsymbol{R}_{x,k} \boldsymbol{\Theta}_\perp \boldsymbol{\xi}_k + 2\,\text{Re}\Big\{\boldsymbol{u}^*\boldsymbol{\Theta}^* \sum_{k=1}^N \boldsymbol{R}_{x,k}\boldsymbol{\Theta}_\perp \boldsymbol{\xi}_k\Big\} \\ &- 2\,\text{Re}\Big\{\sum_{k=1}^N \boldsymbol{p}_{dx,k}^* \boldsymbol{\Theta}\boldsymbol{u}\Big\} - 2\,\text{Re}\Big\{\sum_{k=1}^N \boldsymbol{p}_{dx,k}^* \boldsymbol{\Theta}_\perp \boldsymbol{\xi}_k\Big\}\end{aligned} \tag{10}$$

where $\boldsymbol{R}_{x,k} = \mathbb{E}\{\boldsymbol{x}_{k,n}^* \boldsymbol{x}_{k,n}\}$ is the covariance matrix of $\boldsymbol{x}_{k,n}$, and $\boldsymbol{p}_{dx,k} = \mathbb{E}\{d_k(n)\boldsymbol{x}_{k,n}^*\}$ is the covariance vector between the input data $\boldsymbol{x}_{k,n}$ and the reference output data $d_k(n)$.

Lemma 1: Problem (8) has a unique solution with respect to $\boldsymbol{u}$ and $\{\boldsymbol{\epsilon}_k\}_{k=1}^N$ if the perturbations $\{\boldsymbol{\epsilon}_k\}_{k=1}^N$ lie in a subspace orthogonal to span($\boldsymbol{\Theta}$). ■

Proof of Lemma 1 is provided in Appendix A. We shall now derive a distributed algorithm to seek the minimizer of (8). Focusing on the terms that depend on $\boldsymbol{u}$ in (10), and setting parameters $\boldsymbol{\xi}_k$ to their optimum values $\boldsymbol{\xi}_k^o$, we consider first the global cost function over the variable $\boldsymbol{u}$:

$$\begin{aligned}J_u^{\text{glob}}(\boldsymbol{u}) &= \sum_{k=1}^N \Big(\boldsymbol{u}^*\boldsymbol{\Theta}^*\boldsymbol{R}_{x,k}\boldsymbol{\Theta}\boldsymbol{u} + 2\,\text{Re}\Big\{\boldsymbol{u}^*\boldsymbol{\Theta}^*\boldsymbol{R}_{x,k}\boldsymbol{\Theta}_\perp\boldsymbol{\xi}_k^o\Big\} \\ &\quad - 2\,\text{Re}\Big\{\boldsymbol{p}_{dx,k}^*\boldsymbol{\Theta}\boldsymbol{u}\Big\} + g_k(\boldsymbol{\xi}_k^o)\Big) \\ &= \sum_{k=1}^N J_{u,k}(\boldsymbol{u})\end{aligned} \tag{11}$$

where $g_k(\boldsymbol{\xi}_k^o)$ collects all the terms depending only on $\boldsymbol{\xi}_k^o$ in (10). The term $\sum_{k=1}^N \mathbb{E}\{|d_k(n)|^2\}$ is discarded because it is constant with respect to the arguments $\boldsymbol{u}$ and $\{\boldsymbol{\xi}_k\}_{k=1}^N$. Since $J_u^{\text{glob}}(\boldsymbol{u})$ has a unique minimizer for all nodes over the network, we can use a single-task adapt-then-combine (ATC) diffusion strategy to estimate $\boldsymbol{u}^o$ [13], [15]. We introduce a right-stochastic matrix $\boldsymbol{C}$ with nonnegative entries $c_{\ell k}$ such that:

$$\sum_{k=1}^N c_{\ell k} = 1, \quad \text{and} \quad c_{\ell k} = 0 \text{ if } k \notin \mathcal{N}_\ell \tag{12}$$

With each node k, we associate the local cost over the variable $\boldsymbol{u}$:

$$J_{u,k}^{\text{loc}}(\boldsymbol{u}) = \sum_{\ell \in \mathcal{N}_k} c_{\ell k} J_{u,\ell}(\boldsymbol{u}) \tag{13}$$

Observe that $\sum_{k=1}^N J_{u,k}^{\text{loc}}(\boldsymbol{u}) = J_u^{\text{glob}}(\boldsymbol{u})$ because matrix $\boldsymbol{C}$ is right-stochastic. Since $J_u^{\text{glob}}(\boldsymbol{u})$ is quadratic with respect to $\boldsymbol{u}$, it can be expressed at each node k as follows:

$$\begin{aligned}J_u^{\text{glob}}(\boldsymbol{u}) &= J_{u,k}^{\text{loc}}(\boldsymbol{u}) + \sum_{\ell \neq k} J_{\boldsymbol{u},\ell}^{\text{loc}}(\boldsymbol{u}) \\ &= J_{u,k}^{\text{loc}}(\boldsymbol{u}) + \sum_{\ell \neq k} \|\boldsymbol{u} - \boldsymbol{u}^o\|^2_{\nabla^2 J_{u,\ell}^{\text{loc}}}\end{aligned} \tag{14}$$

where $\nabla^2 J_{u,\ell}^{\text{loc}}$ denotes the Hessian matrix of $J_{u,\ell}^{\text{loc}}(\boldsymbol{u})$ with respect to $\boldsymbol{u}$, and $\|\boldsymbol{u}\|_{\boldsymbol{\Sigma}}^2$ is the squared norm of $\boldsymbol{u}$ weighted by any positive semi-definite matrix $\boldsymbol{\Sigma}$, i.e., $\|\boldsymbol{u}\|_{\boldsymbol{\Sigma}}^2 = \boldsymbol{u}^*\boldsymbol{\Sigma}\boldsymbol{u}$. Following an argument based on the Rayleigh-Ritz characterization of eigenvalues [13, Sec. 3.1], we approximate $\nabla^2 J_{\boldsymbol{u},\ell}^{\text{loc}}$ by a multiple of the identity matrix, so that $\|\boldsymbol{u} - \boldsymbol{u}^o\|^2_{\nabla^2 J_{\boldsymbol{u},\ell}^{\text{loc}}} \approx b_{\ell k}\|\boldsymbol{u} - \boldsymbol{u}^o\|^2$.

Minimizing (14) in two successive steps yields:

$$\boldsymbol{\phi}_{k,n} = \boldsymbol{u}_{k,n-1} - \mu \nabla J_{u,k}^{\text{loc}}(\boldsymbol{u}_{k,n-1}) \tag{15}$$

$$\boldsymbol{u}_{k,n} = \boldsymbol{\phi}_{k,n} + \mu \sum_{\ell \neq k} b_{\ell k}(\boldsymbol{u}^o - \boldsymbol{u}_{k,n-1}) \tag{16}$$

where μ is a positive step size. Its choice to ensure stability of the algorithm will be elaborated on later in Sec. IV. Now, note the following. First, iteration (16) requires knowledge of $\boldsymbol{u}^o$, which is not available. Each node ℓ has a readily available, however, an approximation for $\boldsymbol{u}^o$, which is $\boldsymbol{\phi}_{k,n}$. Therefore, we replace $\boldsymbol{u}^o$ by $\boldsymbol{\phi}_{k,n}$ in (16). Second, $\boldsymbol{\phi}_{k,n}$ at node k is generally a better estimate for $\boldsymbol{u}^o$ than $\boldsymbol{u}_{k,n-1}$ since it is obtained by incorporating information from the neighbors through (15). Therefore, we replace $\boldsymbol{u}_{k,n-1}$ by $\boldsymbol{\phi}_{k,n}$ in (16). Then, absorbing coefficients $b_{\ell k}$ into another set of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2017.2671789, IEEE Journal of Selected Topics in Signal Processing

nonnegative coefficients that satisfies:

$$\sum_{\ell=1}^{N} a_{\ell k} = 1, \quad \text{and} \quad a_{\ell k} = 0 \text{ if } \ell \notin \mathcal{N}_k, \tag{17}$$

which means that matrix $\boldsymbol{A}$ with entries $a_{\ell k}$ is left-stochastic, using an instantaneous approximation of the gradient, and limiting the summation in (16) to the neighbors of node ℓ (see [13], [15] for more details on a similar derivation in the context of single-task diffusion strategies), we can update $\boldsymbol{u}_{k,n}$ as follows:

$$\boldsymbol{\phi}_{k,n} = \boldsymbol{u}_{k,n-1} + \mu \sum_{\ell \in \mathcal{N}_k} c_{\ell k} \boldsymbol{\Theta}^* \boldsymbol{x}_{\ell,n}^* \big[d_\ell(n) - \boldsymbol{x}_{\ell,n}(\boldsymbol{\Theta}\boldsymbol{u}_{k,n-1}) - \boldsymbol{x}_{\ell,n}(\boldsymbol{\Theta}_\perp \boldsymbol{\xi}_{\ell,n-1}) \big] \tag{18}$$

$$\boldsymbol{u}_{k,n} = \sum_{\ell \in \mathcal{N}_k} a_{\ell k} \boldsymbol{\phi}_{\ell,n} \tag{19}$$

where $\boldsymbol{\xi}_{k,n-1}$ is an estimate for the unknown minimizer $\boldsymbol{\xi}_k^o$, to be evaluated as explained further ahead in (21).

Focusing on the terms that depend on $\{\boldsymbol{\xi}_k\}_{k=1}^N$ in (10), and setting parameter $\boldsymbol{u}$ to its optimum value $\boldsymbol{u}^o$, we consider the global cost function over the variables $\boldsymbol{\xi}_k$:

$$\begin{aligned} & J_\xi^{\text{glob}}\big(\{\boldsymbol{\xi}_k\}_{k=1}^N\big) \\ &= \sum_{k=1}^{N} \Big(\boldsymbol{\xi}_k^* \boldsymbol{\Theta}_\perp^* \boldsymbol{R}_{x,k} \boldsymbol{\Theta}_\perp \boldsymbol{\xi}_k + 2\,\text{Re}\Big\{ \boldsymbol{\xi}_k^* \boldsymbol{\Theta}_\perp^* \boldsymbol{R}_{x,k} \boldsymbol{\Theta} \boldsymbol{u}^o \Big\} \\ &\quad - 2\,\text{Re}\Big\{ \boldsymbol{p}_{dx,k}^* \boldsymbol{\Theta}_\perp \boldsymbol{\xi}_k \Big\} \Big) + g_k'(\boldsymbol{u}^o) \\ &= \sum_{k=1}^{N} J_{\xi,k}(\boldsymbol{\xi}_k) \end{aligned} \tag{20}$$

where $g_k'(\boldsymbol{\xi}_k^o)$ collects all the terms depending only on $\boldsymbol{u}^o$ in (10). Now since the parameters $\boldsymbol{\xi}_k$ are node-specific, if no further constraints are imposed, they can be updated independently of each other via an LMS-type update:

$$\begin{aligned} \boldsymbol{\xi}_{k,n} &= \boldsymbol{\xi}_{k,n-1} \\ &\quad + \mu \boldsymbol{\Theta}_\perp^* \boldsymbol{x}_{k,n}^* \big[d_k(n) - \boldsymbol{x}_{k,n}(\boldsymbol{\Theta}\boldsymbol{u}_{k,n-1} + \boldsymbol{\Theta}_\perp \boldsymbol{\xi}_{k,n-1}) \big] \end{aligned} \tag{21}$$

At each time instant n, node k updates its parameters $\boldsymbol{u}_{k,n-1}$ and $\boldsymbol{\xi}_{k,n-1}$ using (18)–(19) and (21), respectively. The local estimate $\boldsymbol{w}_{k,n}$ is then given by:

$$\boldsymbol{w}_{k,n} = \boldsymbol{\Theta}\boldsymbol{u}_{k,n} + \boldsymbol{\Theta}_\perp \boldsymbol{\xi}_{k,n} \tag{22}$$

It is interesting to note that we can rewrite the algorithm without using the auxiliary variables $\boldsymbol{u}_{k,n}$ and $\{\boldsymbol{\xi}_{k,n}\}_{k=1}^N$, by substituting the relations:

$$\boldsymbol{u}_{k,n} = (\boldsymbol{\Theta}^*\boldsymbol{\Theta})^{-1}\boldsymbol{\Theta}^* \boldsymbol{w}_{k,n} \tag{23}$$

$$\boldsymbol{\xi}_{k,n} = (\boldsymbol{\Theta}_\perp^*\boldsymbol{\Theta}_\perp)^{-1}\boldsymbol{\Theta}_\perp^* \boldsymbol{w}_{k,n} \tag{24}$$

into (18)–(19) and (21), respectively. Selecting $\boldsymbol{C} = \boldsymbol{I}_N$ to avoid exchanging raw data and node-specific components, we can implement the update of $\boldsymbol{w}_{k,n-1}$ to an intermediate value $\boldsymbol{\psi}_{k,n}$ as follows:

$$\begin{aligned} \boldsymbol{\psi}_{k,n} &\overset{(a)}{=} \boldsymbol{\Theta}\boldsymbol{\phi}_{k,n} + \boldsymbol{\Theta}_\perp \boldsymbol{\xi}_{k,n} \\ &\overset{(b)}{=} \boldsymbol{\Theta}\boldsymbol{u}_{k,n-1} + \boldsymbol{\Theta}_\perp \boldsymbol{\xi}_{k,n-1} + \mu \Big[(\boldsymbol{\Theta}\boldsymbol{\Theta}^* + \boldsymbol{\Theta}_\perp \boldsymbol{\Theta}_\perp^*) \\ &\quad \times \boldsymbol{x}_{k,n}^* [d_k(n) - \boldsymbol{x}_{k,n}(\boldsymbol{\Theta}\boldsymbol{u}_{k,n-1} + \boldsymbol{\Theta}_\perp \boldsymbol{\xi}_{k,n-1})] \Big] \\ &= \boldsymbol{w}_{k,n-1} + \mu \boldsymbol{S}_{\boldsymbol{\Theta}} \boldsymbol{x}_{k,n}^* \big[d_k(n) - \boldsymbol{x}_{k,n} \boldsymbol{w}_{k,n-1} \big] \end{aligned} \tag{25}$$

with $\boldsymbol{S}_{\boldsymbol{\Theta}} = \boldsymbol{\Theta}\boldsymbol{\Theta}^* + \boldsymbol{\Theta}_\perp \boldsymbol{\Theta}_\perp^*$. For step (a), we use (22) with the intermediate value $\boldsymbol{\phi}_{k,n}$ of $\boldsymbol{u}_{k,n}$ in (18) and $\boldsymbol{\xi}_{k,n}$. Step (b) follows from their adaptation steps (18) and (21). Now substituting (19) in (22) to aggregate the intermediate estimates of $\boldsymbol{u}_{k,n}$ from the neighbors of node k, we arrive at the combination step:

$$\begin{aligned} \boldsymbol{w}_{k,n} &\overset{(22)}{=} \boldsymbol{\Theta}\boldsymbol{u}_{k,n} + \boldsymbol{\Theta}_\perp \boldsymbol{\xi}_{k,n} \\ &\overset{(c)}{=} \boldsymbol{\Theta} \sum_{\ell \in \mathcal{N}_k} a_{\ell k} (\boldsymbol{\Theta}^*\boldsymbol{\Theta})^{-1} \boldsymbol{\Theta}^* \boldsymbol{\psi}_{k,n} + \boldsymbol{\Theta}_\perp (\boldsymbol{\Theta}_\perp^* \boldsymbol{\Theta}_\perp)^{-1} \boldsymbol{\Theta}_\perp^* \boldsymbol{\psi}_{k,n} \\ &= \sum_{\ell \in \mathcal{N}_k} a_{\ell k} \boldsymbol{P}_{\boldsymbol{\Theta}} \boldsymbol{\psi}_{\ell,n} + \boldsymbol{P}_{\boldsymbol{\Theta}_\perp} \boldsymbol{\psi}_{k,n} \end{aligned} \tag{26}$$

where $\boldsymbol{P}_{\boldsymbol{\Theta}} = \boldsymbol{\Theta}(\boldsymbol{\Theta}^*\boldsymbol{\Theta})^{-1}\boldsymbol{\Theta}^*$ and $\boldsymbol{P}_{\boldsymbol{\Theta}_\perp} = \boldsymbol{I}_L - \boldsymbol{P}_{\boldsymbol{\Theta}}$ are the projection matrices over subspaces $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}_\perp$. For step (c), we use (23)–(24) with the intermediate estimate $\boldsymbol{\psi}_{k,n}$. Finally, we arrive at the ATC strategy summarized in Algorithm 1.

The first step in (28) is an adaptation step where node k uses the data realizations $\{d_k(n), \boldsymbol{x}_{k,n}\}$ to update its existing estimate $\boldsymbol{w}_{k,n-1}$ to an intermediate value $\boldsymbol{\psi}_{k,n}$. All other nodes in the network are performing a similar step. The second step in (29) is an aggregation step. To update its intermediate estimate to $\boldsymbol{w}_{k,n}$, each node k combines the existing estimates of its neighbors in the common latent subspace $\boldsymbol{\Theta}$ to build up a common representation, and refines it with a node-specific value in $\boldsymbol{\Theta}_\perp$. In the special case when $\boldsymbol{A} = \boldsymbol{I}_N$, so that no information exchange is performed, the ATC strategy reduces to a non-cooperative solution where each node k runs its own individual descent algorithm.

Matrix $\boldsymbol{S}_{\boldsymbol{\Theta}}$ in the adaptation step (28) is positive-definite. It arises from the calculation of the gradient of (10) with respect to $\boldsymbol{u}$ and $\boldsymbol{\xi}_k$. The algorithm can be simplified by replacing $\boldsymbol{S}_{\boldsymbol{\Theta}}$ by $\boldsymbol{I}_L$ in (28) without compromising the convergence of the method (as analyzed further ahead in Section IV). We then arrive at the recursion:

$$\boldsymbol{\psi}_{k,n} = \boldsymbol{w}_{k,n-1} + \mu\, \boldsymbol{x}_{k,n}^* \big[d_k(n) - \boldsymbol{x}_{k,n} \boldsymbol{w}_{k,n-1} \big] \tag{27}$$

Strictly speaking, observe that $\boldsymbol{S}_{\boldsymbol{\Theta}} = \boldsymbol{I}_L$ if, and only if, the columns of $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}_\perp$ form an orthonormal basis of $\mathbb{R}^L$. Note that the adaptation step (27) is the LMS solution for minimizing the cost in (10) with respect to $\boldsymbol{w}_k$.

Before leaving this section, we would like to point out that the algorithm described in [31], which addresses direct models by stacking global and local variables in an augmented parameter vector, may be used to solve problem (8), provided that an appropriate variable change is performed in order to make the latent variables $\boldsymbol{u}_{k,n}$ and $\boldsymbol{\xi}_{k,n}$ explicit in $\boldsymbol{w}_{k,n}$. The resulting algorithm has the same performance as Algorithm 1 defined by (28), (29), but, obviously, they do not have the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2017.2671789, IEEE Journal of Selected Topics in Signal Processing

Algorithm 1: ATC diffusion LMS with node-specific subspace constraints

Parameters: Preset
- positive step-size μ for all agents;
- left-stochastic combination matrix $\boldsymbol{A}$;
- full-rank matrix $\boldsymbol{\Theta}$ with columns $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$.

Initialization: Set initial weights $\boldsymbol{w}_{k,0} = \boldsymbol{0}$ for all $k \in \{1, ..., N\}$.

Algorithm: At each time instant $n \geq 1$, and for each agent k, update $\boldsymbol{w}_{k,n}$ as:

$$\boldsymbol{\psi}_{k,n} = \boldsymbol{w}_{k,n-1} + \mu \boldsymbol{S}_{\boldsymbol{\Theta}}\, \boldsymbol{x}_{k,n}^{*} \big[d_k(n) - \boldsymbol{x}_{k,n} \boldsymbol{w}_{k,n-1} \big] \tag{28}$$

$$\boldsymbol{w}_{k,n} = \sum_{\ell \in \mathcal{N}_k} a_{\ell k} \boldsymbol{P}_{\boldsymbol{\Theta}} \boldsymbol{\psi}_{\ell,n} + \boldsymbol{P}_{\boldsymbol{\Theta}_\perp} \boldsymbol{\psi}_{k,n} \tag{29}$$

same form since they do not operate in the same domain. This structural difference has a major consequence for Algorithm 1. As already explained, it can be further tuned by replacing the matrix $\boldsymbol{S}_{\boldsymbol{\Theta}}$ in (28) by any positive definite matrix while ensuring convergence of the method. This extra degree of freedom will be taken into account in the analysis of the algorithm, where the only condition on $\boldsymbol{S}_{\boldsymbol{\Theta}}$ is to be positive definite. We will also show that setting $\boldsymbol{S}_{\boldsymbol{\Theta}}$ to $\boldsymbol{I}_L$, besides simplifying Algorithm 1, can greatly improve its performance.

B. Node-specific subspace constraints with norm-bounded projections

The second formulation we consider is to relax the constraint that node-specific components $\{\boldsymbol{\epsilon}_k\}_{k=1}^N$ must lie in span$(\boldsymbol{\Theta}_\perp)$. We now assume that they are norm-bounded in some sense. The problem is formulated as follows:

$$\begin{aligned} &\min_{\boldsymbol{u}, \{\boldsymbol{\epsilon}_k\}_{k=1}^N} J^{\text{glob}}\big(\boldsymbol{u}, \{\boldsymbol{\epsilon}_k\}_{k=1}^N\big) \\ &\text{subject to } \sum_{k=1}^N \|\boldsymbol{P}_{\boldsymbol{\Theta}} \boldsymbol{\epsilon}_k\|^2 \leq \nu_1, \quad \sum_{k=1}^N \|\boldsymbol{P}_{\boldsymbol{\Theta}_\perp} \boldsymbol{\epsilon}_k\|^2 \leq \nu_2 \end{aligned} \tag{30}$$

Since the objective function and the constraints are convex in $(\boldsymbol{u}, \{\boldsymbol{\epsilon}_k\}_{k=1}^N)$, the constrained problem (30) can be formulated as a regularized optimization problem that consists of minimizing a global cost of the form [49]:

$$\begin{aligned} J^{\text{glob}}\big(\boldsymbol{u}, \{\boldsymbol{\epsilon}_k\}_{k=1}^N\big) &= \sum_{k=1}^N \mathbb{E}\big\{ |d_k(n) - \boldsymbol{x}_{k,n}(\boldsymbol{\Theta}\boldsymbol{u} + \boldsymbol{\epsilon}_k)|^2 \big\} \\ &\quad + \eta_1 \sum_{k=1}^N \|\boldsymbol{P}_{\boldsymbol{\Theta}} \boldsymbol{\epsilon}_k\|^2 + \eta_2 \sum_{k=1}^N \|\boldsymbol{P}_{\boldsymbol{\Theta}_\perp} \boldsymbol{\epsilon}_k\|^2 \end{aligned} \tag{31}$$

where η_1 and η_2 are positive regularization parameters that are related to the bounds ν_1 and ν_2.

Lemma 2: Problem (30) has a unique solution with respect to $\boldsymbol{u}$ and $\{\boldsymbol{\epsilon}_k\}_{k=1}^N$. ■

Proof of Lemma 2 is provided in Appendix B. Other norms such as the general $\ell_{p,q}$-norm may be used with $\boldsymbol{\epsilon}_k$ in (30), depending on the application. Some form of regularization on $\boldsymbol{u}$ may also be included. However, using the ℓ_2-norm with $\boldsymbol{\epsilon}_k$ in (30) enables us to solve the problem with respect to $\boldsymbol{w}_k$, without using the auxiliary variables $\boldsymbol{u}$ and $\{\boldsymbol{\epsilon}_k\}_{k=1}^N$. Indeed, let us rewrite (31) as follows:

$$\begin{aligned} J^{\text{glob}}\big(\boldsymbol{u}, \{\boldsymbol{w}_k\}_{k=1}^N\big) &= \sum_{k=1}^N \mathbb{E}\big\{ |d_k(n) - \boldsymbol{x}_{k,n} \boldsymbol{w}_k|^2 \big\} \\ &\quad + \eta_1 \sum_{k=1}^N \|\boldsymbol{P}_{\boldsymbol{\Theta}}(\boldsymbol{w}_k - \boldsymbol{\Theta}\boldsymbol{u})\|^2 + \eta_2 \sum_{k=1}^N \|\boldsymbol{P}_{\boldsymbol{\Theta}_\perp} \boldsymbol{w}_k\|^2 \end{aligned} \tag{32}$$

The optimality condition relative to $\boldsymbol{u}$ gives:

$$\sum_{k=1}^N \boldsymbol{\Theta}^* \boldsymbol{P}_{\boldsymbol{\Theta}} (\boldsymbol{w}_k^o - \boldsymbol{\Theta}\boldsymbol{u}^o) = \boldsymbol{0} \tag{33}$$

from which the optimal parameter vector $\boldsymbol{u}^o$ can be expressed as:

$$\boldsymbol{u}^o = \frac{1}{N} \sum_{k=1}^N (\boldsymbol{\Theta}^* \boldsymbol{\Theta})^{-1} \boldsymbol{\Theta}^* \boldsymbol{w}_k^o \tag{34}$$

Substituting (34) into (32), and using that $\boldsymbol{P}_{\boldsymbol{\Theta}}$ is Hermitian and idempotent (i.e., $\boldsymbol{P}_{\boldsymbol{\Theta}} = \boldsymbol{P}_{\boldsymbol{\Theta}}^2$), yields:

$$\begin{aligned} J^{\text{glob}}\big(\{\boldsymbol{w}_k\}_{k=1}^N\big) &= \sum_{k=1}^N \mathbb{E}\big\{ |d_k(n) - \boldsymbol{x}_{k,n} \boldsymbol{w}_k|^2 \big\} \\ &\quad + \eta_1 \sum_{k=1}^N \Big\| \boldsymbol{P}_{\boldsymbol{\Theta}} \boldsymbol{w}_k - \frac{1}{N} \sum_{\ell=1}^N \boldsymbol{P}_{\boldsymbol{\Theta}} \boldsymbol{w}_\ell \Big\|^2 + \eta_2 \sum_{k=1}^N \|\boldsymbol{P}_{\boldsymbol{\Theta}_\perp} \boldsymbol{w}_k\|^2 \end{aligned} \tag{35}$$

Node k can apply a steepest-descent iteration to minimize the cost in (35) with respect to $\{\boldsymbol{w}_k\}_{k=1}^N$. Computing the gradient vector of (35) we get:

$$\begin{aligned} \nabla J^{\text{glob}} = \Big[&(\boldsymbol{R}_{x,k} \boldsymbol{w}_k - \boldsymbol{p}_{dx}) + \eta_1 \big(\boldsymbol{P}_{\boldsymbol{\Theta}} \boldsymbol{w}_k - \frac{1}{N} \sum_{\ell=1}^N \boldsymbol{P}_{\boldsymbol{\Theta}} \boldsymbol{w}_\ell \big) \\ &+ \eta_2 \boldsymbol{P}_{\boldsymbol{\Theta}_\perp} \boldsymbol{w}_k \Big]^* \end{aligned} \tag{36}$$

Starting from an initial condition $\boldsymbol{w}_{k,0}$, we arrive at the steepest descent iteration:

$$\begin{aligned} \boldsymbol{w}_{k,n} = \boldsymbol{w}_{k,n-1} &- \mu \Big[(\boldsymbol{R}_{x,k} \boldsymbol{w}_{k,n-1} - \boldsymbol{p}_{dx}) + \eta_2 \boldsymbol{P}_{\boldsymbol{\Theta}_\perp} \boldsymbol{w}_{k,n-1} \Big] \\ &- \mu\eta_1 \Big(\boldsymbol{P}_{\boldsymbol{\Theta}} \boldsymbol{w}_{k,n-1} - \frac{1}{N} \sum_{\ell=1}^N \boldsymbol{P}_{\boldsymbol{\Theta}} \boldsymbol{w}_{\ell,n-1} \Big) \end{aligned} \tag{37}$$

This iteration indicates that the update term involves adding two correction terms to $\boldsymbol{w}_{k,n-1}$. Among many other forms, we can implement the update in two successive steps by adding one correction term at a time:

$$\begin{aligned} \boldsymbol{\psi}_{k,n} = \boldsymbol{w}_{k,n-1} &- \mu \Big[(\boldsymbol{R}_{x,k} \boldsymbol{w}_{k,n-1} - \boldsymbol{p}_{dx}) + \eta_2 \boldsymbol{P}_{\boldsymbol{\Theta}_\perp} \boldsymbol{w}_{k,n-1} \Big] \end{aligned} \tag{38}$$

$$\begin{aligned} \boldsymbol{w}_{k,n} = \boldsymbol{\psi}_{k,n} &- \mu\eta_1 \Big(\boldsymbol{P}_{\boldsymbol{\Theta}} \boldsymbol{w}_{k,n-1} - \frac{1}{N} \sum_{\ell=1}^N \boldsymbol{P}_{\boldsymbol{\Theta}} \boldsymbol{w}_{\ell,n-1} \Big) \end{aligned} \tag{39}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2017.2671789, IEEE Journal of Selected Topics in Signal Processing

Step (38) updates $\boldsymbol{w}_{k,n-1}$ to an intermediate value $\boldsymbol{\psi}_{k,n}$. We now revise (38)–(39) to achieve a diffusion LMS type algorithm. The intermediate value $\boldsymbol{\psi}_{\ell,n}$ at node ℓ is generally expected to be a better estimate for $\boldsymbol{w}_\ell^o$ than $\boldsymbol{w}_{\ell,n-1}$ since it is updated by the first step (38). Therefore, we replace $\boldsymbol{w}_{\ell,n-1}$ by $\boldsymbol{\psi}_{\ell,n}$ in the second step (39) as follows to get:

$$\begin{aligned}\boldsymbol{w}_{k,n} &= \boldsymbol{\psi}_{k,n} - \mu\eta_1\Big(\boldsymbol{P}_{\boldsymbol{\Theta}}\boldsymbol{\psi}_{k,n} - \frac{1}{N}\sum_{\ell=1}^{N}\boldsymbol{P}_{\boldsymbol{\Theta}}\boldsymbol{\psi}_{\ell,n}\Big)\\ &= (\boldsymbol{\psi}_{k,n} - \boldsymbol{P}_{\Theta}\boldsymbol{\psi}_{k,n}) + \Big((1-\mu\eta_1)\boldsymbol{P}_{\Theta}\boldsymbol{\psi}_{k,n} + \sum_{\ell=1}^{N}\frac{\mu\eta_1}{N}\,\boldsymbol{\psi}_{\ell,n}\Big)\end{aligned} \tag{40}$$

Observe that $\boldsymbol{P}_{\boldsymbol{\Theta}_\perp}\boldsymbol{\psi}_{k,n} = \boldsymbol{\psi}_{k,n} - \boldsymbol{P}_{\boldsymbol{\Theta}}\,\boldsymbol{\psi}_{k,n}$, and introduce the coefficients $a_{\ell k} = \frac{\mu\eta_1}{N}$ for $\ell \neq k$, and $a_{kk} = 1 - \mu\eta_1 + \frac{\mu\eta_1}{N}$. We get:

$$\boldsymbol{w}_{k,n} = \boldsymbol{P}_{\boldsymbol{\Theta}_\perp}\boldsymbol{\psi}_{k,n} + \sum_{\ell=1}^{N} a_{\ell k}\boldsymbol{P}_{\boldsymbol{\Theta}}\boldsymbol{\psi}_{\ell,n} \tag{41}$$

Considering that each node in the network can only share information with its neighbors, and using instantaneous approximations for $\boldsymbol{R}_{x,k}$ and $\boldsymbol{p}_{dx}$, we arrive at:

$$\begin{aligned}\boldsymbol{\psi}_{k,n} &= (\boldsymbol{I}_L - \mu\eta_2\boldsymbol{P}_{\boldsymbol{\Theta}_\perp})\boldsymbol{w}_{k,n-1}\\ &\quad + \mu\boldsymbol{x}_{k,n}^*\big[d_k(n) - \boldsymbol{x}_{k,n}\boldsymbol{w}_{k,n-1}\big]\end{aligned} \tag{42}$$

$$\boldsymbol{w}_{k,n} = \sum_{\ell\in\mathcal{N}_k} a_{\ell k}\boldsymbol{P}_{\boldsymbol{\Theta}}\boldsymbol{\psi}_{\ell,n} + \boldsymbol{P}_{\boldsymbol{\Theta}_\perp}\boldsymbol{\psi}_{k,n} \tag{43}$$

with $a_{kk} = 1 - \mu\eta_1 + \frac{\mu\eta_1}{|\mathcal{N}_k|}$ and $a_{\ell k} = \frac{\mu\eta_1}{|\mathcal{N}_k|}$ for $\ell \in \mathcal{N}_k$ and $\ell \neq k$. Note that, for sufficiently small step-sizes μ_k, these coefficients are nonnegative and satisfy $\sum_{\ell=1}^{N} a_{\ell k} = 1$ for all k. We will treat these coefficients as free parameters that can be chosen by the designer according to these conditions (i.e., nonnegative coefficients that add up to one on each column of matrix $\boldsymbol{A}$). We summarize this statement in Algorithm 2.

Algorithms 1 and 2 employ the same aggregation step in (29) and (45). Node k combines the intermediate estimates of its neighbors in the common subspace $\boldsymbol{\Theta}$ without affecting the local contribution in the complementary subspace $\boldsymbol{\Theta}_\perp$. The norm constraint (30) in $\boldsymbol{\Theta}_\perp$ leads to a leaky-LMS alike term in the adaptation step (44).

Let us now examine two special cases of Algorithm 2. First, in the case where $\boldsymbol{\Theta} = \boldsymbol{0}$, problem (31) reduces to a regularized least-mean squares problem with $\boldsymbol{w}_k = \boldsymbol{\epsilon}_k$. That is, the algorithm reduces to the non-cooperative leaky-LMS algorithm. On the other hand, if $\boldsymbol{\Theta} = \boldsymbol{I}_L$, the algorithm reduces to diffusion LMS.

Before leaving this section, we briefly discuss the complexity of Algorithms 1 and 2. Both algorithms have the same adapt-then-combine structure as the diffusion LMS except that each node needs to project data on $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}_\perp$. This means that each node k only needs to update the $L \times 1$ parameter vectors $\boldsymbol{\psi}_{k,n}$ and $\boldsymbol{w}_{k,n}$ at each time instant. Next, each node k needs to transmit $\boldsymbol{w}_{k,n}$ to its $|\mathcal{N}_k| - 1$ neighbors. A projection performed by a matrix-vector product has a computational complexity of $\mathcal{O}(L\log_2 L)$ [50]. All the other operations performed by each node have a complexity of $\mathcal{O}(L)$.

Algorithm 2: ATC diffusion LMS with node-specific subspace constraints (norm-bounded projections)

Parameters: Preset
- positive step-size μ for all agents;
- full-rank matrix $\boldsymbol{\Theta}$ with columns $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$.

Initialization: Set initial weights $\boldsymbol{w}_{k,0} = \boldsymbol{0}$ for all $k = 1, ..., N$.

Algorithm: For each instant $n \geq 1$, and for each agent k, update $\boldsymbol{w}_{k,n-1}$:

$$\begin{aligned}\boldsymbol{\psi}_{k,n} &= (\boldsymbol{I}_L - \mu\eta_2\boldsymbol{P}_{\boldsymbol{\Theta}_\perp})\boldsymbol{w}_{k,n-1}\\ &\quad + \mu\boldsymbol{x}_{k,n}^*\big[d_k(n) - \boldsymbol{x}_{k,n}\boldsymbol{w}_{k,n-1}\big]\end{aligned} \tag{44}$$

$$\boldsymbol{w}_{k,n} = \sum_{\ell\in\mathcal{N}_k} a_{\ell k}\boldsymbol{P}_{\boldsymbol{\Theta}}\boldsymbol{\psi}_{\ell,n} + \boldsymbol{P}_{\boldsymbol{\Theta}_\perp}\boldsymbol{\psi}_{k,n} \tag{45}$$

IV. Performance and convergence analyses

In this section, we examine the convergence properties and network performance of the proposed adaptive strategies. We shall first describe a convergence framework for a family of distributed algorithms, where Algorithms 1 and 2 are special cases. Quantities specifically related to Algorithms 1 or 2 will be distinguished by superscripts $^{(1)}$ and $^{(2)}$, respectively.

In order to perform the analysis, we collect information from across the network into block vectors and matrices. Let us denote by $\boldsymbol{w}_n$ and $\boldsymbol{w}^o$ the block weight vector at instant n and the block optimum weight vector, both of size $LN \times 1$, that is

$$\boldsymbol{w}_n = \mathrm{col}\{\boldsymbol{w}_{1,n}, \ldots, \boldsymbol{w}_{N,n}\} \tag{46}$$

$$\boldsymbol{w}^o = \mathrm{col}\{\boldsymbol{w}_1^o, \ldots, \boldsymbol{w}_N^o\} \tag{47}$$

We denote the difference between the optimum $\boldsymbol{w}_k^o$ and the instantaneous estimate $\boldsymbol{w}_{k,n}$ by:

$$\boldsymbol{v}_{k,n} = \boldsymbol{w}_k^o - \boldsymbol{w}_{k,n} \tag{48}$$

We collect the weight error vectors $\boldsymbol{v}_{k,n}$ from across all nodes into the block weight error vector:

$$\boldsymbol{v}_n = \mathrm{col}\{\boldsymbol{v}_{1,n}, \ldots, \boldsymbol{v}_{N,n}\} \tag{49}$$

Assumption 1: (Independent inputs) The regression vectors $\boldsymbol{x}_{k,n}$ arise from a stationary random process that is temporally stationary, white, and independent over space with $\boldsymbol{R}_{x,k} = \mathbb{E}\{\boldsymbol{x}_{k,n}^*\boldsymbol{x}_{k,n}\} > 0$. A direct consequence of this condition is that $\boldsymbol{x}_{k,n}$ is independent of $\boldsymbol{v}_{\ell,m}$ for all ℓ and $m \leq n$.

A. Mean weight behavior analysis

The estimation error in (28) and (44) can be rewritten as a function of $\boldsymbol{v}_{k,n}$:

$$d_k(n) - \boldsymbol{x}_{k,n}\boldsymbol{w}_{k,n-1} = z_k(n) + \boldsymbol{x}_{k,n}\boldsymbol{v}_{k,n-1} \tag{50}$$

In what follows, we first show that the weight error update relations for both Algorithms 1 and 2 are of the form:

$$\boldsymbol{v}_n = \boldsymbol{B}_n\boldsymbol{v}_{n-1} - \mu\,\boldsymbol{g}_n - \boldsymbol{r}, \tag{51}$$

with $\boldsymbol{B}_n$ an $LN \times LN$ time-dependent matrix, $\boldsymbol{g}_n$ an $LN \times 1$ zero-mean time-dependent vector, and $\boldsymbol{r}$ a constant $LN \times$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2017.2671789, IEEE Journal of Selected Topics in Signal Processing

1 vector. Consequently, it will be possible to represent their mean weight behavior in the form of a state-transition equation with a bounded driving term:

$$\mathbb{E}\{\boldsymbol{v}_n\} = \boldsymbol{B}\,\mathbb{E}\{\boldsymbol{v}_{n-1}\} - \boldsymbol{r} \tag{52}$$

with $\boldsymbol{B} = \mathbb{E}\{\boldsymbol{B}_{n-1}\}$. Let $\boldsymbol{H}_{x,n}$ be the block diagonal matrix of size $LN \times LN$, and $\boldsymbol{p}_{zx,n}$ the vector of length $LN \times 1$, defined as follows:

$$\boldsymbol{H}_{x,n} \triangleq \text{diag}\{\boldsymbol{x}_{1,n}^*\boldsymbol{x}_{1,n},\, \ldots,\, \boldsymbol{x}_{N,n}^*\boldsymbol{x}_{N,n}\} \tag{53}$$

$$\boldsymbol{p}_{zx,n} \triangleq \text{col}\{z_1(n)\boldsymbol{x}_{1,n}^*,\, \ldots,\, z_N(n)\boldsymbol{x}_{N,n}^*\} \tag{54}$$

The expectation of $\boldsymbol{H}_{x,n}$ and $\boldsymbol{p}_{xz,n}$ are given by:

$$\boldsymbol{H}_x \triangleq \mathbb{E}\{\boldsymbol{H}_{x,n}\} = \text{diag}\,\{\boldsymbol{R}_{x,1},\, \ldots, \boldsymbol{R}_{N,2}\} \tag{55}$$

$$\boldsymbol{p}_{zx} \triangleq \mathbb{E}\{\boldsymbol{p}_{zx,n}\} = \boldsymbol{0} \tag{56}$$

1) Mean weight behavior of Algorithm 1: Define the intermediate weight error vector $\widetilde{\boldsymbol{\psi}}_{k,n}$:

$$\widetilde{\boldsymbol{\psi}}_{k,n} = \boldsymbol{w}_k^o - \boldsymbol{\psi}_{k,n} \tag{57}$$

and collect these vectors from across all nodes into the block weight error vector:

$$\widetilde{\boldsymbol{\psi}}_n = \text{col}\{\widetilde{\boldsymbol{\psi}}_{1,n}, \ldots, \widetilde{\boldsymbol{\psi}}_{N,n}\} \tag{58}$$

Subtracting $\boldsymbol{w}_k^o$ from both sides of the update relation (28), and using relation (50), leads to the update equation for $\widetilde{\boldsymbol{\psi}}_n$:

$$\widetilde{\boldsymbol{\psi}}_n = (\boldsymbol{I}_{LN} - \mu\boldsymbol{D}_{\boldsymbol{S}_\Theta}\boldsymbol{H}_{x,n})\boldsymbol{v}_{n-1} - \mu\boldsymbol{D}_{\boldsymbol{S}_\Theta}\boldsymbol{p}_{zx,n} \tag{59}$$

where $\boldsymbol{D}_{\boldsymbol{S}_\Theta} = \text{diag}\{\boldsymbol{S}_\Theta, \ldots, \boldsymbol{S}_\Theta\}$ is an $LN \times LN$ block diagonal matrix with $\boldsymbol{S}_\Theta$ as diagonal entries. Let $\boldsymbol{\mathcal{A}} = \boldsymbol{A} \otimes \boldsymbol{I}_L$. Defining $\boldsymbol{D}_{\boldsymbol{P}_\Theta}$ and $\boldsymbol{D}_{\boldsymbol{P}_{\Theta_\perp}}$ as the $LN \times LN$ block diagonal matrices with $\boldsymbol{P}_\Theta$ and $\boldsymbol{P}_{\Theta_\perp}$ as diagonal entries, respectively, equation (29) can be written in vector form as:

$$\boldsymbol{w}_n = \big(\boldsymbol{\mathcal{A}}^\top\boldsymbol{D}_{\boldsymbol{P}_\Theta} + \boldsymbol{D}_{\boldsymbol{P}_{\Theta_\perp}}\big)\psi_n \tag{60}$$

Subtracting $\boldsymbol{w}^o$ from both sides of the above expression, we have:

$$\boldsymbol{v}_n = \big(\boldsymbol{\mathcal{A}}^\top\boldsymbol{D}_{\boldsymbol{P}_\Theta} + \boldsymbol{D}_{\boldsymbol{P}_{\Theta_\perp}}\big)\widetilde{\psi}_n - \big(\boldsymbol{\mathcal{A}}^\top\boldsymbol{D}_{\boldsymbol{P}_\Theta} + \boldsymbol{D}_{\boldsymbol{P}_{\Theta_\perp}} - \boldsymbol{I}_{LN}\big)\boldsymbol{w}^o \tag{61}$$

Combining this equation with (59), the weight error update relation can be written in a single expression:

$$\begin{aligned}\boldsymbol{v}_n = \big(\boldsymbol{\mathcal{A}}^\top\boldsymbol{D}_{\boldsymbol{P}_\Theta} + \boldsymbol{D}_{\boldsymbol{P}_{\Theta_\perp}}\big)\big[(\boldsymbol{I}_{LN} - \mu\boldsymbol{D}_{\boldsymbol{S}_\Theta}\boldsymbol{H}_{x,n})\boldsymbol{v}_{n-1} \\ - \mu\boldsymbol{D}_{\boldsymbol{S}_\Theta}\boldsymbol{p}_{zx,n}\big] - (\boldsymbol{\mathcal{A}}^\top - \boldsymbol{I}_{LN})\boldsymbol{D}_{\boldsymbol{P}_\Theta}\boldsymbol{w}^o\end{aligned} \tag{62}$$

Now we denote several terms in the weight error expression (62) by:

$$\boldsymbol{B}_n^{(1)} = \big(\boldsymbol{\mathcal{A}}^\top\boldsymbol{D}_{\boldsymbol{P}_\Theta} + \boldsymbol{D}_{\boldsymbol{P}_{\Theta_\perp}}\big)(\boldsymbol{I}_{LN} - \mu\boldsymbol{D}_{\boldsymbol{S}_\Theta}\boldsymbol{H}_{x,n}) \tag{63}$$

$$\boldsymbol{g}_n^{(1)} = \big(\boldsymbol{\mathcal{A}}^\top\boldsymbol{D}_{\boldsymbol{P}_\Theta} + \boldsymbol{D}_{\boldsymbol{P}_{\Theta_\perp}}\big)\boldsymbol{D}_{\boldsymbol{S}_\Theta}\boldsymbol{p}_{zx,n} \tag{64}$$

$$\boldsymbol{r}^{(1)} = (\boldsymbol{\mathcal{A}}^\top - \boldsymbol{I}_{LN})\boldsymbol{D}_{\boldsymbol{P}_\Theta}\boldsymbol{w}^o, \tag{65}$$

and the associated expected values:

$$\begin{aligned}\boldsymbol{B}^{(1)} &\triangleq \mathbb{E}\{\boldsymbol{B}_n^{(1)}\} \\ &= \big(\boldsymbol{\mathcal{A}}^\top\boldsymbol{D}_{\boldsymbol{P}_\Theta} + \boldsymbol{D}_{\boldsymbol{P}_{\Theta_\perp}}\big)(\boldsymbol{I}_{LN} - \mu\boldsymbol{D}_{\boldsymbol{S}_\Theta}\boldsymbol{H}_x)\end{aligned} \tag{66}$$

$$\boldsymbol{g}^{(1)} \triangleq \mathbb{E}\{\boldsymbol{g}_n^{(1)}\} = \boldsymbol{0} \tag{67}$$

With the above notation, the weight error update relation (62) can be written as:

$$\boldsymbol{v}_n = \boldsymbol{B}_n^{(1)}\,\boldsymbol{v}_{n-1} - \mu\,\boldsymbol{g}_n^{(1)} - \boldsymbol{r}^{(1)} \tag{68}$$

Taking the expectation on both sides of (68), and using Assumption 1, we arrive at the mean weight behavior for Algorithm 1:

$$\mathbb{E}\{\boldsymbol{v}_n\} = \boldsymbol{B}^{(1)}\,\mathbb{E}\{\boldsymbol{v}_{n-1}\} - \boldsymbol{r}^{(1)} \tag{69}$$

2) Mean weight behavior of Algorithm 2: Subtracting $\boldsymbol{w}_k^o$ from both sides of the update relation (44), and using relation (50), yields:

$$\begin{aligned}\widetilde{\boldsymbol{\psi}}_n = (\boldsymbol{I} - \mu\eta_2\boldsymbol{D}_{\boldsymbol{P}_{\boldsymbol{\Theta}_\perp}} - \mu\boldsymbol{H}_{x,n})\boldsymbol{v}_{n-1} \\ - \mu(\boldsymbol{p}_{zx,n} - \eta_2\boldsymbol{D}_{\boldsymbol{P}_{\boldsymbol{\Theta}_\perp}}\boldsymbol{w}^o)\end{aligned} \tag{70}$$

Subtracting $\boldsymbol{w}^o$ from both sides of (45), we have:

$$\boldsymbol{v}_n = \big(\boldsymbol{\mathcal{A}}^\top\boldsymbol{D}_{\boldsymbol{P}_\Theta} + \boldsymbol{D}_{\boldsymbol{P}_{\Theta_\perp}}\big)\widetilde{\boldsymbol{\psi}}_n - \big(\boldsymbol{\mathcal{A}}^\top\boldsymbol{D}_{\boldsymbol{P}_\Theta} + \boldsymbol{D}_{\boldsymbol{P}_{\Theta_\perp}} - \boldsymbol{I}_{LN}\big)\boldsymbol{w}^o \tag{71}$$

Combining this equation with (70), the weight error update relation can be written in a single expression:

$$\begin{aligned}\boldsymbol{v}_n = \big(\boldsymbol{\mathcal{A}}^\top\boldsymbol{D}_{\boldsymbol{P}_\Theta} + \boldsymbol{D}_{\boldsymbol{P}_{\Theta_\perp}}\big)\big[(\boldsymbol{I}_{LN} - \mu\eta_2\boldsymbol{D}_{\boldsymbol{P}_{\boldsymbol{\Theta}_\perp}} - \mu\boldsymbol{H}_{x,n})\boldsymbol{v}_{n-1} \\ - \mu(\boldsymbol{p}_{zx,n} - \eta_2\boldsymbol{D}_{\boldsymbol{P}_{\boldsymbol{\Theta}_\perp}}\boldsymbol{w}^o)\big] - (\boldsymbol{\mathcal{A}}^\top - \boldsymbol{I}_{LN})\boldsymbol{D}_{\boldsymbol{P}_\Theta}\,\boldsymbol{w}^o\end{aligned} \tag{72}$$

where we used the fact that $\boldsymbol{I}_{LN} = \boldsymbol{D}_{\boldsymbol{P}_\Theta} + \boldsymbol{D}_{\boldsymbol{P}_{\Theta_\perp}}$. Next, we denote several terms in the weight error expression (72) by:

$$\boldsymbol{B}_n^{(2)} = \big(\boldsymbol{\mathcal{A}}^\top\boldsymbol{D}_{\boldsymbol{P}_\Theta} + \boldsymbol{D}_{\boldsymbol{P}_{\Theta_\perp}}\big)(\boldsymbol{I}_{LN} - \mu\eta_2\boldsymbol{D}_{\boldsymbol{P}_{\boldsymbol{\Theta}_\perp}} - \mu\boldsymbol{H}_{x,n}) \tag{73}$$

$$\boldsymbol{g}_n^{(2)} = \big(\boldsymbol{\mathcal{A}}^\top\boldsymbol{D}_{\boldsymbol{P}_\Theta} + \boldsymbol{D}_{\boldsymbol{P}_{\Theta_\perp}}\big)\boldsymbol{p}_{zx,n} \tag{74}$$

$$\begin{aligned}\boldsymbol{r}^{(2)} = (\boldsymbol{\mathcal{A}}^\top - \boldsymbol{I}_{LN})\boldsymbol{D}_{\boldsymbol{P}_\Theta}\boldsymbol{w}^o \\ - \mu\eta_2\big(\boldsymbol{\mathcal{A}}^\top\boldsymbol{D}_{\boldsymbol{P}_\Theta} + \boldsymbol{D}_{\boldsymbol{P}_{\Theta_\perp}}\big)\boldsymbol{D}_{\boldsymbol{P}_{\boldsymbol{\Theta}_\perp}}\boldsymbol{w}^o\end{aligned} \tag{75}$$

and the associated expected values:

$$\begin{aligned}\boldsymbol{B}^{(2)} &\triangleq \mathbb{E}\{\boldsymbol{B}_n^{(2)}\} \\ &= \big(\boldsymbol{\mathcal{A}}^\top\boldsymbol{D}_{\boldsymbol{P}_\Theta} + \boldsymbol{D}_{\boldsymbol{P}_{\Theta_\perp}}\big)(\boldsymbol{I}_{LN} - \mu\eta_2\boldsymbol{D}_{\boldsymbol{P}_{\boldsymbol{\Theta}_\perp}} - \mu\boldsymbol{H}_x)\end{aligned} \tag{76}$$

$$\boldsymbol{g}^{(2)} \triangleq \mathbb{E}\{\boldsymbol{g}_n^{(2)}\} = \boldsymbol{0} \tag{77}$$

With the above notation, the weight error update relation (72) can be written as:

$$\boldsymbol{v}_n = \boldsymbol{B}_n^{(2)}\,\boldsymbol{v}_{n-1} - \mu\,\boldsymbol{g}_n^{(2)} - \boldsymbol{r}^{(2)} \tag{78}$$

Taking the expectation on both sides of (78), and using Assumption 1, we get the mean weight behavior of Algorithm 2:

$$\mathbb{E}\{\boldsymbol{v}_n\} = \boldsymbol{B}^{(2)}\,\mathbb{E}\{\boldsymbol{v}_{n-1}\} - \boldsymbol{r}^{(2)} \tag{79}$$

3) Stability in the mean: The mean-weight error recursions (69) and (79) are of the same form as (52). The convergence of such recursive state-transition equations, with a bounded driving term, is determined by the stability of matrix $\boldsymbol{B}$. Algorithm parameters should be chosen to satisfy the mean stability condition $\rho(\boldsymbol{B}) < 1$, where $\rho(\cdot)$ denotes spectral radius of its matrix argument. In this case, the bias of the algorithm will be given by:

$$\lim_{n\to\infty}\mathbb{E}\{\boldsymbol{v}_n\} = -(\boldsymbol{I}_{LN} - \boldsymbol{B})^{-1}\boldsymbol{r} \tag{80}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2017.2671789, IEEE Journal of Selected Topics in Signal Processing

We shall now establish two results that provide ranges for selecting the step size μ to ensure convergence in the mean for each algorithm.

Theorem 1: (Stability in the mean for Algorithm 1) Assume data model (1) and Assumption 1 hold. We select a doubly stochastic matrix $\boldsymbol{A}$. Assume $\{\boldsymbol{\Theta}, \boldsymbol{\Theta}_\perp\}$ forms an orthonormal basis of $\mathbb{R}^L$. Then, for any initial condition, Algorithm 1 asymptotically converges in the mean if the step-size satisfies:

$$0 < \mu < \frac{2}{\max_k \lambda_{\max}(\boldsymbol{R}_{x,k})} \tag{81}$$

where $\lambda_{\max}(\cdot)$ denotes the maximum eigenvalue of its matrix argument.

Proof: The convergence of (69) is determined by the stability of matrix $\boldsymbol{B}^{(1)}$. The required mean stability condition is met by selecting μ so that:

$$\rho\big((\boldsymbol{\mathcal{A}}^\top \boldsymbol{D}_{\boldsymbol{P}_\Theta} + \boldsymbol{D}_{\boldsymbol{P}_{\Theta_\perp}})(\boldsymbol{I}_{LN} - \mu \boldsymbol{S}_\Theta \boldsymbol{H}_x)\big) < 1 \tag{82}$$

Let $\boldsymbol{x} = \text{col}\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ be any block vector of size $LN \times 1$. We have:

$$\begin{aligned}
&\big\|(\boldsymbol{\mathcal{A}}^\top \boldsymbol{D}_{\boldsymbol{P}_\Theta} + \boldsymbol{D}_{\boldsymbol{P}_{\Theta_\perp}})\boldsymbol{x}\big\|^2 \\
&\quad = \sum_{i=1}^N \Big\| \sum_{j=1}^N a_{ji}\, \boldsymbol{P}_{\boldsymbol{\Theta}} \boldsymbol{x}_j + \boldsymbol{P}_{\boldsymbol{\Theta}_\perp} \boldsymbol{x}_i \Big\|^2 \qquad (83) \\
&\quad = \sum_{i=1}^N \Big(\Big\| \sum_{j=1}^N a_{ji}\, \boldsymbol{P}_{\boldsymbol{\Theta}} \boldsymbol{x}_j \Big\|^2 + \big\| \boldsymbol{P}_{\boldsymbol{\Theta}_\perp} \boldsymbol{x}_i \big\|^2 \Big) \qquad (84)
\end{aligned}$$

Given that $\boldsymbol{A}$ is left stochastic, namely, $\sum_{j=1}^N a_{ji} = 1$ with $a_{ji} \geq 0$, Jensen's inequality guarantees:

$$\Big\| \sum_{j=1}^N a_{ji}\, \boldsymbol{P}_{\boldsymbol{\Theta}} \boldsymbol{x}_j \Big\|^2 \leq \sum_{j=1}^N a_{ji} \big\| \boldsymbol{P}_{\boldsymbol{\Theta}} \boldsymbol{x}_j \big\|^2 \tag{85}$$

Consequently, the quantity in (84) can be upper-bounded as follows:

$$\begin{aligned}
&\sum_{i=1}^N \Big(\Big\| \sum_{j=1}^N a_{ji}\, \boldsymbol{P}_{\boldsymbol{\Theta}} \boldsymbol{x}_j \Big\|^2 + \big\| \boldsymbol{P}_{\boldsymbol{\Theta}_\perp} \boldsymbol{x}_i \big\|^2 \Big) \\
&\leq \sum_{i=1}^N \sum_{j=1}^N a_{ji} \big\| \boldsymbol{P}_{\boldsymbol{\Theta}} \boldsymbol{x}_j \big\|^2 + \sum_{i=1}^N \big\| \boldsymbol{P}_{\boldsymbol{\Theta}_\perp} \boldsymbol{x}_i \big\|^2 \qquad (86) \\
&\overset{(a)}{=} \sum_{j=1}^N \big\| \boldsymbol{P}_{\boldsymbol{\Theta}} \boldsymbol{x}_j \big\|^2 + \sum_{i=1}^N \big\| \boldsymbol{P}_{\boldsymbol{\Theta}_\perp} \boldsymbol{x}_i \big\|^2 \qquad (87) \\
&= \|\boldsymbol{x}\|^2 \qquad (88)
\end{aligned}$$

where for step (a) we use that $\boldsymbol{A}$ is right stochastic, namely, $\sum_{i=1}^N a_{ji} = 1$. We conclude that:

$$\big\| \boldsymbol{\mathcal{A}}^\top \boldsymbol{D}_{\boldsymbol{P}_\Theta} + \boldsymbol{D}_{\boldsymbol{P}_{\Theta_\perp}} \big\| \leq 1 \tag{89}$$

We know that the spectral radius of any matrix $\boldsymbol{X}$ satisfies $\rho(\boldsymbol{X}) \leq \|\boldsymbol{X}\|$, for any induced norm. Then we have:

$$\begin{aligned}
&\rho\big((\boldsymbol{\mathcal{A}}^\top \boldsymbol{D}_{\boldsymbol{P}_\Theta} + \boldsymbol{D}_{\boldsymbol{P}_{\Theta_\perp}})(\boldsymbol{I}_{LN} - \mu \boldsymbol{S}_\Theta \boldsymbol{H}_x)\big) \\
&\leq \big\| \boldsymbol{\mathcal{A}}^\top \boldsymbol{D}_{\boldsymbol{P}_\Theta} + \boldsymbol{D}_{\boldsymbol{P}_{\Theta_\perp}} \big\| \, \big\| \boldsymbol{I}_{LN} - \mu \boldsymbol{S}_\Theta \boldsymbol{H}_x \big\| \qquad (90) \\
&\overset{(89)}{\leq} \big\| \boldsymbol{I}_{LN} - \mu \boldsymbol{S}_\Theta \boldsymbol{H}_x \big\| \qquad (91)
\end{aligned}$$

The mean stability condition is thus met by selecting μ so that: $\big\| \boldsymbol{I}_{LN} - \mu \boldsymbol{S}_\Theta \boldsymbol{H}_x \big\| < 1$. In the case where $\{\boldsymbol{\Theta}, \boldsymbol{\Theta}_\perp\}$ forms an orthonormal basis of $\mathbb{R}^L$, then $\boldsymbol{S}_\Theta = \boldsymbol{I}_L$. This leads us to the condition in (81). ■

Theorem 2: (Stability in the mean for Algorithm 2) Assume data model (1) and Assumption 1 hold. We select a doubly stochastic matrix $\boldsymbol{A}$. Then, for any initial condition, Algorithm 2 asymptotically converges in the mean if the step-size satisfies:

$$0 < \mu < \frac{2}{\max_k \lambda_{\max}(\eta_2 \boldsymbol{P}_{\boldsymbol{\Theta}_\perp} + \boldsymbol{R}_{x,k})} \tag{92}$$

Proof: The convergence of (79) is determined by the stability of matrix $\boldsymbol{B}^{(2)}$. Considering that:

$$\begin{aligned}
&\rho\big((\boldsymbol{\mathcal{A}}^\top \boldsymbol{D}_{\boldsymbol{P}_\Theta} + \boldsymbol{D}_{\boldsymbol{P}_{\Theta_\perp}})(\boldsymbol{I}_{LN} - \mu\eta_2\, \boldsymbol{D}_{\boldsymbol{P}_{\Theta_\perp}} - \mu \boldsymbol{H}_x)\big) \\
&\leq \big\| \boldsymbol{I}_{LN} - \mu\eta_2\, \boldsymbol{D}_{\boldsymbol{P}_{\Theta_\perp}} - \mu \boldsymbol{H}_x \big\|
\end{aligned} \tag{93}$$

since $\|\boldsymbol{\mathcal{A}}^\top \boldsymbol{D}_{\boldsymbol{P}_\Theta} + \boldsymbol{D}_{\boldsymbol{P}_{\Theta_\perp}}\| \leq 1$, the mean stability condition is met by selecting μ so that $\big\| \boldsymbol{I}_{LN} - \mu\eta_2\, \boldsymbol{D}_{\boldsymbol{P}_{\Theta_\perp}} - \mu\, \boldsymbol{H}_x \big\| < 1$. This leads us to the condition in (92). Furthermore, by Weyl's theorem, we have $\lambda_{\max}(\eta_2 \boldsymbol{P}_{\boldsymbol{\Theta}_\perp} + \boldsymbol{R}_{x,k}) \leq \eta_2 + \lambda_{\max}(\boldsymbol{R}_{x,k})$ since $\boldsymbol{P}_{\boldsymbol{\Theta}_\perp}$ and $\boldsymbol{R}_{x,k}$ are Hermitian matrices and $\lambda_{\max}(\boldsymbol{P}_{\boldsymbol{\Theta}_\perp}) = 1$. This leads to the sufficient condition:

$$0 < \mu < \frac{2}{\eta_2 + \max_k \lambda_{\max}(\boldsymbol{R}_{x,k})} \tag{94}$$

■

B. Mean-square error behavior analysis

We now study the mean-square error behavior of Algorithms 1 and 2. To this end, we consider the general update relation (52) since both algorithms are of this form. From (51), the squared norm $\|\boldsymbol{v}_n\|_{\boldsymbol{\Sigma}}^2$ of the weight vector $\boldsymbol{v}_n$ weighted by any positive semi-definite matrix $\boldsymbol{\Sigma}$, i.e., $\|\boldsymbol{v}_n\|_{\boldsymbol{\Sigma}}^2 = \boldsymbol{v}_n^* \boldsymbol{\Sigma} \boldsymbol{v}_n$, satisfies the following relation:

$$\begin{aligned}
\|\boldsymbol{v}_n\|_{\boldsymbol{\Sigma}}^2 &= \|\boldsymbol{v}_{n-1}\|_{\boldsymbol{B}_n^* \boldsymbol{\Sigma} \boldsymbol{B}_n}^2 - \mu^2 \|\boldsymbol{g}_n\|_{\boldsymbol{\Sigma}} + \|\boldsymbol{r}\|_{\boldsymbol{\Sigma}}^2 \\
&\quad - 2\,\text{Re}\{\boldsymbol{r}^* \boldsymbol{\Sigma} \boldsymbol{B}_n \boldsymbol{v}_{n-1}\} - 2\mu\,\text{Re}\{\boldsymbol{g}_n^* \boldsymbol{\Sigma}(\boldsymbol{B}_n \boldsymbol{v}_{n-1} - \boldsymbol{r})\}
\end{aligned} \tag{95}$$

Under the independence assumption, and considering that $\boldsymbol{g}_n$ includes the zero-mean noise term z_n which is independent of any other signal, taking expectations of both sides of (95) leads to:

$$\begin{aligned}
\mathbb{E}\{\|\boldsymbol{v}_n\|_{\boldsymbol{\Sigma}}^2\} &= \mathbb{E}\{\|\boldsymbol{v}_{n-1}\|_{\boldsymbol{\Sigma}'}^2\} + \mu^2\, \text{trace}\{\boldsymbol{\Sigma} \mathbb{E}\{\boldsymbol{g}_n \boldsymbol{g}_n^*\}\} \\
&\quad + \|\boldsymbol{r}\|_{\boldsymbol{\Sigma}}^2 - 2\,\text{Re}\big\{\mathbb{E}\{\boldsymbol{r}^* \boldsymbol{\Sigma} \boldsymbol{B}_n \boldsymbol{v}_{n-1}\}\big\}
\end{aligned} \tag{96}$$

In the above expression, $\boldsymbol{\Sigma}$ is any positive semi-definite matrix that the user is free to choose in order to derive different performance metrics, and $\boldsymbol{\Sigma}' = \mathbb{E}\{\boldsymbol{B}_n^* \boldsymbol{\Sigma} \boldsymbol{B}_n\}$. Let $\boldsymbol{G}$ be the expected value of $\mathbb{E}\{\boldsymbol{g}_n \boldsymbol{g}_n^*\}$ in the second term on the RHS of (96). For the two presented algorithms, $\boldsymbol{G}$ is respectively given by:

$$\begin{aligned}
\boldsymbol{G}^{(1)} &= (\boldsymbol{\mathcal{A}}^\top \boldsymbol{D}_{\boldsymbol{P}_\Theta} + \boldsymbol{D}_{\boldsymbol{P}_{\Theta_\perp}}) \boldsymbol{D}_{\boldsymbol{S}_\Theta} \text{diag}\{\sigma_{z,1}^2 \boldsymbol{R}_{x,1}, \\
&\qquad \ldots, \sigma_{z,N}^2 \boldsymbol{R}_{x,N}\}\, \boldsymbol{D}_{\boldsymbol{S}_\Theta}^* (\boldsymbol{\mathcal{A}}^\top \boldsymbol{D}_{\boldsymbol{P}_\Theta} + \boldsymbol{D}_{\boldsymbol{P}_{\Theta_\perp}})^* \qquad (97) \\
\boldsymbol{G}^{(2)} &= \text{diag}\{\sigma_{z,1}^2 \boldsymbol{R}_{x,1}, \ldots, \sigma_{z,N}^2 \boldsymbol{R}_{x,N}\} \qquad (98)
\end{aligned}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2017.2671789, IEEE Journal of Selected Topics in Signal Processing

With $\boldsymbol{G}$, equation (96) is expressed as:

$$\mathbb{E}\{\|\boldsymbol{v}_n\|^2_{\boldsymbol{\Sigma}}\} = \mathbb{E}\{\|\boldsymbol{v}_{n-1}\|^2_{\boldsymbol{\Sigma}'}\} + \mu^2 \text{trace}\,\{\boldsymbol{\Sigma}\boldsymbol{G}\} + \|\boldsymbol{r}\|^2_{\boldsymbol{\Sigma}} - 2\,\text{Re}\big\{\boldsymbol{r}^*\boldsymbol{\Sigma}\,\boldsymbol{B}\,\mathbb{E}\{\boldsymbol{v}_{n-1}\}\big\} \quad (99)$$

Vectorizing matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}'$ by $\boldsymbol{\sigma} = \text{vec}(\boldsymbol{\Sigma})$ and $\boldsymbol{\sigma}' = \text{vec}(\boldsymbol{\Sigma}')$, it can be verified that:

$$\boldsymbol{\sigma}' = \boldsymbol{K}\boldsymbol{\sigma} \quad (100)$$

where the $(LN)^2 \times (LN)^2$ matrix $\boldsymbol{K}$ is given by:

$$\boldsymbol{K} = \mathbb{E}\{\boldsymbol{B}_n^\top \otimes \boldsymbol{B}_n^*\} \approx \boldsymbol{B}^\top \otimes \boldsymbol{B}^* \quad (101)$$

The above approximation can be used provided that the step size is sufficiently small so that the influence of the second-degree term in μ can be neglected [13]. Equation (99) can then be expressed as:

$$\mathbb{E}\{\|\boldsymbol{v}_n\|^2_{\boldsymbol{\sigma}}\} = \mathbb{E}\{\|\boldsymbol{v}_{n-1}\|^2_{\boldsymbol{K}\boldsymbol{\sigma}}\} + \boldsymbol{s}_{n-1}^\top\,\boldsymbol{\sigma} \quad (102)$$

where we use the notation $\|\cdot\|_\Sigma$ and $\|\cdot\|_\sigma$ interchangeably, and

$$\boldsymbol{s}_{n-1} = \text{vec}\Big(\mu^2\,\boldsymbol{G} + \boldsymbol{r}\boldsymbol{r}^* - 2\,\text{Re}\big\{\boldsymbol{B}\,\mathbb{E}\{\boldsymbol{v}_{n-1}\}\boldsymbol{r}^*\big\}\Big) \quad (103)$$

Theorem 3: (Mean-square stability) Assume data model (1) and Assumption 1 hold. Assume further that the step size μ is sufficiently small to guarantee the stability in the mean of the algorithms, and to ensure that (102) can be used as a reasonable representation for the evolution of the (weighted) mean-square error. Mean-square stability of cooperative algorithms characterized by (51) requires the step-size μ to be chosen such that it ensures the stability of matrix $\boldsymbol{K}$ (in addition to the mean stability condition $\rho(\boldsymbol{B}) < 1$).

Proof: Iterating (102) starting from $n = 0$, we find that

$$\mathbb{E}\{\|\boldsymbol{v}_n\|^2_{\boldsymbol{\sigma}}\} = \|\boldsymbol{v}_0\|^2_{\boldsymbol{K}^n\boldsymbol{\sigma}} + \sum_{i=1}^{n} \boldsymbol{s}_{n-i}^\top \boldsymbol{K}^{i-1}\boldsymbol{\sigma} \quad (104)$$

with initial condition $\boldsymbol{v}_0 = \boldsymbol{w}^o - \boldsymbol{w}_0$. Provided that $\boldsymbol{K}$ is stable, the first term on the RHS of (104) converges to zero as $n \to \infty$. We know from (52) that $\mathbb{E}\{\boldsymbol{v}_n\}$ is bounded because (52) is a BIBO stable recursion with a bounded driving term $\boldsymbol{r}$. The second term on the RHS of (104) then converges as $n \to \infty$. We conclude that $\mathbb{E}\{\|\boldsymbol{v}_n\|^2_{\boldsymbol{\sigma}}\}$ converges to a bounded value as $n \to \infty$, and the algorithm is mean-square stable. ■

Theorem 4: (Transient MSD) Consider a sufficiently small step size μ to ensure mean and mean-square stabilities. The MSD learning curve $\zeta_n = \frac{1}{N}\mathbb{E}\{\|\boldsymbol{v}_n\|^2\}$ of the cooperative algorithms characterized by (51), obtained by setting $\boldsymbol{\Sigma} = \frac{1}{N}\boldsymbol{I}_{LN}$, evolves according to the following recursion for $n \geq 1$:

$$\zeta_n = \zeta_{n-1} + \frac{1}{N}\big[(\boldsymbol{\gamma}_{n-1} + \boldsymbol{s}_{n-1})^\top \text{vec}(\boldsymbol{I}_{LN}) - \|\boldsymbol{v}_0\|^2_{(\boldsymbol{I}_{(LN)^2} - \boldsymbol{K})\boldsymbol{K}^{n-1}\boldsymbol{\sigma}}\big] \quad (105)$$

$$\boldsymbol{\gamma}_n = \boldsymbol{K}^\top\boldsymbol{\gamma}_{n-1} + (\boldsymbol{K} - \boldsymbol{I}_{(LN)^2})^\top \boldsymbol{s}_{n-1} \quad (106)$$

with initial conditions $\zeta_0 = \frac{1}{N}\|\boldsymbol{v}_0\|^2$ and $\boldsymbol{\gamma}_0 = \boldsymbol{0}$.

Proof: Comparing (104) at instants n and $n-1$, we can relate $\mathbb{E}\{\|\boldsymbol{v}_n\|^2_{\boldsymbol{\sigma}}\}$ and $\mathbb{E}\{\|\boldsymbol{v}_{n-1}\|^2_{\boldsymbol{\sigma}}\}$ as follows:

$$\begin{aligned}\mathbb{E}\{\|\boldsymbol{v}_n\|^2_{\boldsymbol{\sigma}}\} &= \mathbb{E}\{\|\boldsymbol{v}_{n-1}\|^2_{\boldsymbol{\sigma}}\} - \|\boldsymbol{v}_0\|^2_{(\boldsymbol{I}_{(LN)^2} - \boldsymbol{K})\boldsymbol{K}^{n-1}\boldsymbol{\sigma}} \\ &\quad + \sum_{i=1}^{n} \boldsymbol{s}_{n-i}^\top \boldsymbol{K}^{i-1}\boldsymbol{\sigma} - \sum_{i=1}^{n-1} \boldsymbol{s}_{n-1-i}^\top \boldsymbol{K}^{i-1}\boldsymbol{\sigma} \\ &= \mathbb{E}\{\|\boldsymbol{v}_{n-1}\|^2_{\boldsymbol{\sigma}}\} - \|\boldsymbol{v}_0\|^2_{(\boldsymbol{I}_{(LN)^2} - \boldsymbol{K})\boldsymbol{K}^{n-1}\boldsymbol{\sigma}} \\ &\quad + \boldsymbol{s}_{n-1}^\top\boldsymbol{\sigma} + \sum_{i=2}^{n} \boldsymbol{s}_{n-i}^\top \boldsymbol{K}^{i-1}\boldsymbol{\sigma} - \sum_{i=1}^{n-1} \boldsymbol{s}_{n-1-i}^\top \boldsymbol{K}^{i-1}\boldsymbol{\sigma}\end{aligned} \quad (107)$$

Introducing the notation

$$\boldsymbol{\gamma}_{n-1} = \Big[\sum_{i=2}^{n} \boldsymbol{s}_{n-i}^\top \boldsymbol{K}^{i-1} - \sum_{i=1}^{n-1} \boldsymbol{s}_{n-1-i}^\top \boldsymbol{K}^{i-1}\Big]^\top \quad (108)$$

we can reformulate the recursive expression (107) as follows:

$$\mathbb{E}\{\|\boldsymbol{v}_n\|^2_{\boldsymbol{\sigma}}\} = \mathbb{E}\{\|\boldsymbol{v}_{n-1}\|^2_{\boldsymbol{\sigma}}\} - \|\boldsymbol{v}_0\|^2_{(\boldsymbol{I}_{(LN)^2} - \boldsymbol{K})\boldsymbol{K}^{n-1}\boldsymbol{\sigma}} + (\boldsymbol{\gamma}_{n-1} + \boldsymbol{s}_{n-1})^\top\boldsymbol{\sigma} \quad (109)$$

$$\boldsymbol{\gamma}_n = \boldsymbol{K}^\top\boldsymbol{\gamma}_{n-1} + (\boldsymbol{K} - \boldsymbol{I}_{(LN)^2})^\top \boldsymbol{s}_{n-1} \quad (110)$$

with $\boldsymbol{\gamma}_0 = \boldsymbol{0}$. To derive the transient curve for the MSD, replace $\boldsymbol{\sigma}$ by $\frac{1}{N}\text{vec}\{\boldsymbol{I}_{LN}\}$. ■

Corollary 1: (Steady-state MSD) If the step size is chosen sufficiently small to ensure mean and mean-square convergence, then the steady-state MSD, defined as $\zeta_\infty = \lim_{n\to\infty}\zeta_n$, is given by:

$$\zeta_\infty = \frac{1}{N}\,\boldsymbol{s}_\infty^\top\,(\boldsymbol{I}_{(LN)^2} - \boldsymbol{K})^{-1}\text{vec}(\boldsymbol{I}_{LN}) \quad (111)$$

with $\boldsymbol{s}_\infty = \lim_{n\to\infty}\boldsymbol{s}_n$ determined by (103), using $\mathbb{E}\{\boldsymbol{v}_\infty\} = \lim_{n\to\infty}\mathbb{E}\{\boldsymbol{v}_n\}$ determined by (80).

Proof: From expression (102), we get:

$$\lim_{n\to\infty} \mathbb{E}\{\|\boldsymbol{v}_n\|^2_{(\boldsymbol{I}_{(LN)^2} - \boldsymbol{K})\boldsymbol{\sigma}}\} = \boldsymbol{s}_\infty^\top\boldsymbol{\sigma} \quad (112)$$

Observe that the MSD calculation requires us to choose $\boldsymbol{\sigma}$ that satisfies:

$$(\boldsymbol{I}_{(LN)^2} - \boldsymbol{K})\,\boldsymbol{\sigma} = \frac{1}{N}\text{vec}(\boldsymbol{I}_{LN}) \quad (113)$$

This leads to expression (111). ■

V. Simulations

In this section, we report simulation results that illustrate the theoretical results. All agents were initialized with zero parameter vectors $\boldsymbol{w}_{k,0} = \boldsymbol{0}$ for all k. Simulated curves were obtained by averaging over 100 runs as we obtained sufficiently smooth curves to check the consistency with theoretical results.

A. Algorithm validation

We considered a network consisting of 12 agents with interconnections shown in Fig. 1(a). The parameter vectors to be estimated were of length $L = 5$. The input data $\boldsymbol{x}_{k,n}$ were generated from circularly-symmetric zero-mean complex Gaussian distributions. White input data were considered first, by setting:

$$\boldsymbol{R}_{x,k} = \sigma^2_{x,k}\,\boldsymbol{I}_5 \quad (114)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2017.2671789, IEEE Journal of Selected Topics in Signal Processing

Next, correlated input data, characterized by the following covariance matrix, were considered:

$$\boldsymbol{R}_{x,k} = \sigma_{x,k}^2 \times \begin{pmatrix} 1 & -.4+.3j & .2-.1j & .1-.05j & .02+.02j \\ -.4-.3j & 1 & -.4+.3j & .2-.1j & .1-.05j \\ .2+.1j & -.4-.3j & 1 & -.4+.3j & .2-.1j \\ .1+.05j & .2+.1j & -.4-.3j & 1 & -.4+.3j \\ .02-.02j & .1+.05j & .2+.1j & -.4-.3j & 1 \end{pmatrix} \tag{115}$$

with $j = \sqrt{-1}$ the imaginary unit. The modeling noises $z_{k,n}$ were i.i.d. zero-mean circularly-symmetric Gaussian variables, independent of any other signals. The variances $\sigma_{x,k}^2$ and $\sigma_{z,k}^2$ were sampled from $\mathcal{U}(0.8, 1.2)$ and $\mathcal{U}(0.18, 0.22)$, respectively. Their values are shown in Fig. 1(b). We considered two sets of subspace basis vectors. The first set is the standard basis:

$$\boldsymbol{\Theta}_1 = [\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_M], \tag{116}$$

where $\boldsymbol{e}_i$ denotes a vector of length N with 1 at the ith entry and 0 otherwise. Its orthogonal complementary subspace is spanned by $\boldsymbol{\Theta}_{1,\perp} = [\boldsymbol{e}_{M+1}, \ldots, \boldsymbol{e}_L]$. This setup can be interpreted as a variable selection process for information exchange, where the first M entries of the optimal parameter vectors are identical across the network. Parameter M was set to 3. The second set of basis vectors is a complex Vandermonde matrix:

$$\boldsymbol{\Theta}_2 = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ e^{-j\psi_1} & e^{-j\psi_2} & \cdots & e^{-j\psi_M} \\ \cdots & \cdots & \cdots & \cdots \\ e^{-j(L-1)\psi_1} & e^{-j(L-1)\psi_2} & \cdots & e^{-j(L-1)\psi_M} \end{pmatrix} \tag{117}$$

with $\psi_k = \frac{2\pi d}{\lambda_o} \sin\theta_k$. Matrix $\boldsymbol{\Theta}_2$ can represent the array manifold of a uniform linear array (ULA) with inter-element space d, operating at wavelength λ_o with impinging signal directions of angles θ_k. Parameter M was set to 3, with $\theta_1 = \frac{\pi}{6}$, $\theta_2 = \frac{\pi}{4}$, $\theta_3 = \frac{\pi}{3}$ and $d = \frac{\lambda_o}{2}$. We considered three settings to validate the theoretical results.

In the first setting, we assumed that model (6) matches the observation data. The entries of the coefficient vectors $\boldsymbol{u}^o$ and $\boldsymbol{\xi}_k^o$ were sampled from the Gaussian distribution $\mathcal{N}(0, 1)$. The step-size parameter μ for Algorithm 1 was successively set to 0.01 and 0.02. A uniform combination matrix $\boldsymbol{A}$ with $a_{\ell k} = |\mathcal{N}_k|^{-1}$ was used. With $\boldsymbol{\Theta}_1$, note that matrix $\boldsymbol{S}_{\boldsymbol{\Theta}}$ is equal to $\boldsymbol{I}_5$. With $\boldsymbol{\Theta}_2$, it was successively set to $\boldsymbol{\Theta}\boldsymbol{\Theta}^* + \boldsymbol{\Theta}_\perp\boldsymbol{\Theta}_\perp^*$ as in (29), and to $\boldsymbol{I}_5$. The transient behavior and the steady-state MSD were determined theoretically. The results with subspace settings $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$, for white and correlated input data, are shown in Fig. 2. It can be observed that setting $\boldsymbol{S}_{\boldsymbol{\Theta}}$ to $\boldsymbol{I}_5$ for $\boldsymbol{\Theta}_2$ leads to a better convergence behavior. For Algorithm 2, we did not set the parameter η_1 explicitly but we used the same combination matrix $\boldsymbol{A}$ as for Algorithm 1. Parameters (μ, η_2) were set to $(0.02, 0.01)$ with white input data. With correlated input data, the following combinations (μ, η_2) were considered: $\{(0.01, 0.01); (0.01, 0.02); (0.02, 0.01)\}$. The results are shown in Fig. 3. The simulation results match the theoretical results, and illustrate the trade-off between the convergence speed and the steady-state MSD. It can also be observed with Algorithm 2 that a small value for η_2 is preferable since constraining the norm of node-specific components in the complementary subspace $\boldsymbol{\Theta}^\perp$ introduces a bias that can degrade the performance. As leaky-LMS, this kind of regularization can improve the stability of the algorithm for some particular problems and practical applications, at the cost of an extra estimation bias. We then considered another scenario in order to illustrate the interest of the extra degree of freedom provided by η_2 in Algorithm 2. Experimental setups were left unchanged with correlated inputs except for the entries of $\boldsymbol{\xi}_k^o$, which were sampled from Gaussian distribution $\mathcal{N}(0, 0.01)$. We successively set η_2 to 0, 0.1 and 1 in order to progressively constrain the variance of $\boldsymbol{\xi}_k$. Note that with $\eta_2 = 0$, Algorithm 2 reduces to Algorithm 1. The results with $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$ are provided in Fig. 4. The result with non-cooperative LMS is also provided as a reference.

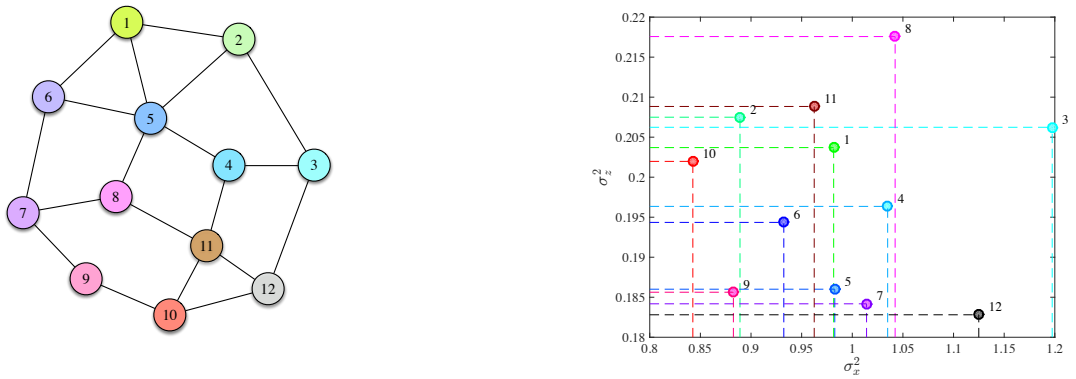
In the second setting, we assumed that the node-specific components $\boldsymbol{\epsilon}_k^o$ in (6) do not strictly lie in the complementary subspace $\boldsymbol{\Theta}^\perp$. To evaluate the robustness of our algorithms and the power of the analytical models, we set:

$$\boldsymbol{\epsilon}_k^o = \boldsymbol{\Theta}\boldsymbol{\nu}_k^o + \boldsymbol{\Theta}_\perp\boldsymbol{\xi}_k^o \tag{118}$$

where $\boldsymbol{\nu}_k^o$ are zero-mean circular Gaussian variables. This setting refers to a non-ideal situation because components $\boldsymbol{\Theta}(\boldsymbol{u}^o + \boldsymbol{\nu}_k^o)$ lie in span$(\boldsymbol{\Theta})$ but differ from one node to another. The entries of $\boldsymbol{u}^o$ and $\boldsymbol{\nu}_k^o$ were sampled from Gaussian distributions $\mathcal{N}(0, 1)$ and $\mathcal{N}(0, 0.01)$, respectively. The step-size μ was set to 0.01 for Algorithms 1 and 2. Parameter η_2 in Algorithm 2 was set to 0.01. Subspace $\boldsymbol{\Theta}_1$ and white input signals were considered to test the model. The transient behavior and the steady-state MSD were determined theoretically. The simulation results provided in Fig. 5 match the theoretical results, and illustrate that cooperation among nodes can still be beneficial when optimal solutions in the subspace $\boldsymbol{\Theta}$ are different but close to each other. This is another illustration of the conclusion reached in [19] for single-task diffusion LMS operating in multitask environments.

In the third setting, we exploited the leaky property of Algorithm 2 to promote its use in real applications. It is well known that the (non-cooperative) leaky LMS algorithm introduces an estimation bias compared to the (non-cooperative) LMS, but improves its robustness when applied to practical applications [51]. In particular, it avoids the so-called weight-drift problem of the LMS algorithm [52]. To highlight this phenomenon in the context of diffusion adaptation, we assumed that, say, the last tap/channel of node #1 was failing to work and was providing consistent null-valued readings, i.e., $[\boldsymbol{x}_{n,1}]_5 = 0$ for all n. We also assumed that, e.g., finite-precision effect was corrupting the combination step (29), or (45), with an additive non-zero mean disturbance $\boldsymbol{q}_k$. The poor conditioning of regressors associated with a non-zero mean disturbance is known to possibly lead to a weight-drift problem. We considered the same experimental setup as in the first experiment with the standard basis $\boldsymbol{\Theta}_1$. We picked each entry of the random vectors $\boldsymbol{q}_k$ according to the Gaussian distribution $\mathcal{N}(10^{-4}, 10^{-8})$. We set η_2 to 0.1. All the vectors $\boldsymbol{w}_k$ were initialized to $\boldsymbol{0}$. Fig. 6 shows the behavior of the weight vector at node #1 for (a) Algorithm 1 with $\boldsymbol{S}_\Theta = \boldsymbol{I}_5$, and (b) Algorithm 2. We can observe the drift of the 5^{th} entry of $\boldsymbol{w}_1$ with Algorithm 1. Algorithm 2 alleviates this effect.

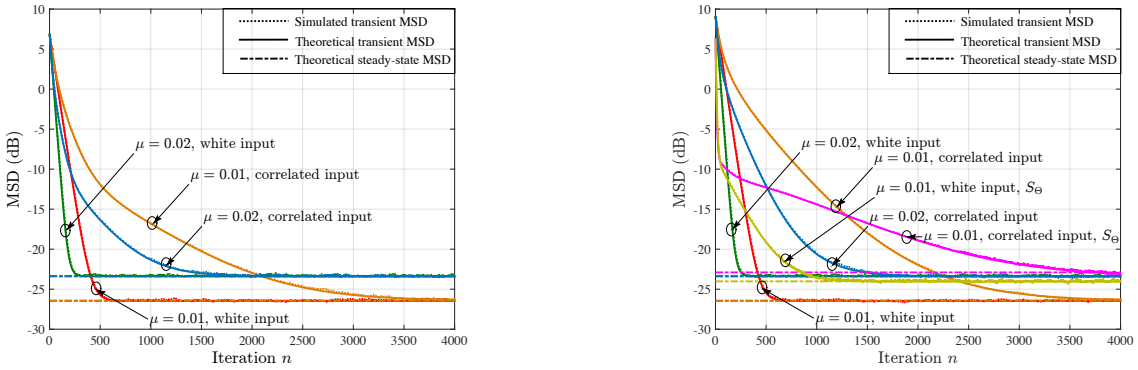
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2017.2671789, IEEE Journal of Selected Topics in Signal Processing



(a) Network topology.

(b) Agent input and noise variances.

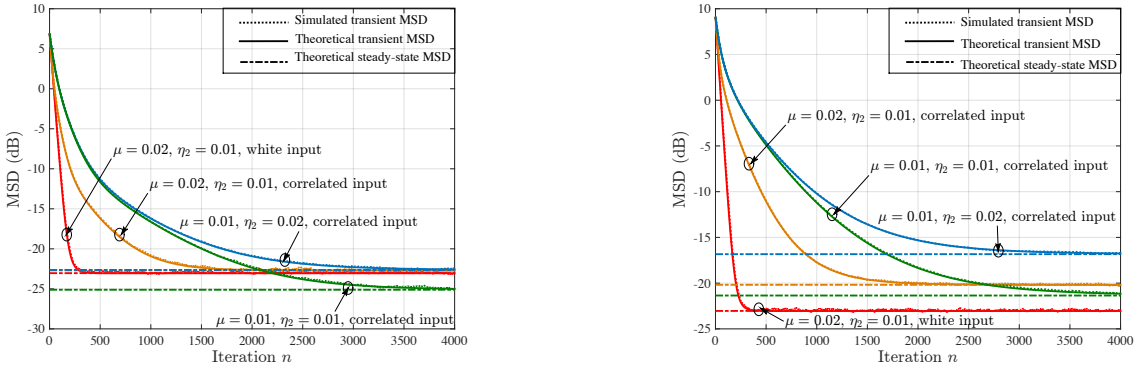
Fig. 1. Network topology and input-noise variances.



(a) Using $\boldsymbol{\Theta}_1$.

(b) Using $\boldsymbol{\Theta}_2$.

Fig. 2. Learning curves and model validation of Algorithm 1 with different settings.



(a) Using $\boldsymbol{\Theta}_1$.

(b) Using $\boldsymbol{\Theta}_2$.

Fig. 3. Learning curves and model validation of Algorithm 2 with different settings.

B. Target localization

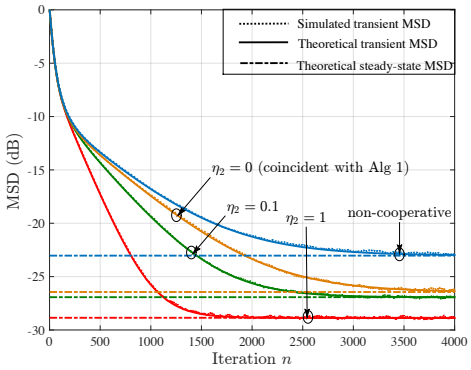
We now consider a target localization problem. Cooperative localization with a diffusion strategy was already addressed in the case of a single target [13], and of multiple nearby targets [4]. We focus here on the case where targets lie in a manifold.

To make the presentation clearer, we assumed that the targets were collinear in $\mathbb{R}^3$. Their locations were estimated by the network with 100 nodes shown in Fig. 7(a). Each node randomly selected a target to localize. Let $\boldsymbol{\mathcal{R}}$ be a member of the rotation group SO(3) defined by the matrix $\boldsymbol{R} = \boldsymbol{R}_x(\theta_x)\,\boldsymbol{R}_y(\theta_y)\,\boldsymbol{R}_z(\theta_z)$, where $\boldsymbol{R}_x(\theta_x)$, $\boldsymbol{R}_y(\theta_y)$ and $\boldsymbol{R}_z(\theta_z)$ are rotation matrices that rotate vectors by an angle of $\theta_{x,y,z}$ around x, y and z axis, respectively. The coordinate vector $\boldsymbol{w}_q^o$ of each target q was generated as follows:

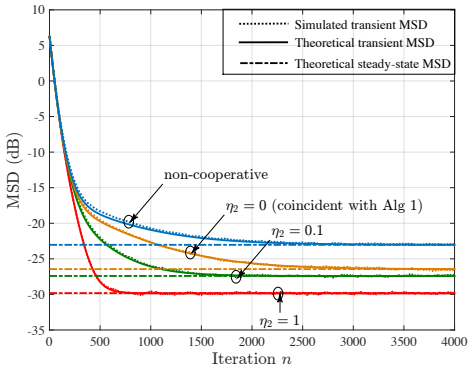
$$\boldsymbol{w}_q^o = \boldsymbol{R}_{1,2}\boldsymbol{u} + \epsilon_q \boldsymbol{r}_3 \tag{119}$$

where $\boldsymbol{R}_{1,2}$ is the matrix composed of the first and second columns of $\boldsymbol{R}$, and $\boldsymbol{r}_3$ corresponds to the third column of $\boldsymbol{R}$. As illustrated in Fig. 7(b), this model means that all targets lie on a common line defined by point $\boldsymbol{R}_{1,2}\boldsymbol{u}$ and direction vector $\boldsymbol{r}_3$. Parameter ϵ_q characterizes the location of each target q on this line. We considered the problem of estimating $\boldsymbol{u}$ (common to all targets) and the parameters ϵ_q

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2017.2671789, IEEE Journal of Selected Topics in Signal Processing



(a) Using $\boldsymbol{\Theta}_1$.



(b) Using $\boldsymbol{\Theta}_2$.

Fig. 4. Learning curves and model validation of the algorithms using $\boldsymbol{\xi}_k$ with small variances.

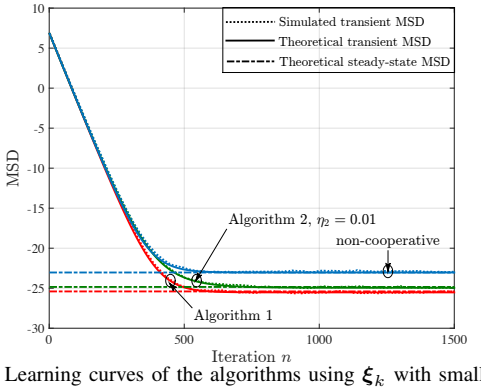


Fig. 5. Learning curves of the algorithms using $\boldsymbol{\xi}_k$ with small variances.

for seven targets. We set the angles and the parameter vectors in (119) as follows:

$$\theta_x = \frac{\pi}{6}, \qquad \theta_y = \frac{\pi}{3}, \qquad \theta_z = \frac{\pi}{4} \tag{120}$$

$$\boldsymbol{v} = [1\ 2]^\top \tag{121}$$

$$\epsilon_1 = 0, \epsilon_2 = 1, \epsilon_3 = 3, \epsilon_4 = 4, \epsilon_5 = 7, \epsilon_6 = 7.5, \epsilon_7 = 9 \tag{122}$$

The distance between each agent k and target q can be expressed in the inner product form:

$$r_{kq} = \boldsymbol{x}_{kq}(\boldsymbol{w}_q^o - \boldsymbol{p}_k) \tag{123}$$

where $\boldsymbol{p}_k$ is the location of agent k, and $\boldsymbol{x}_{kq}$ is the unit-norm row vector pointing from $\boldsymbol{p}_k$ to $\boldsymbol{w}_q^o$. We assumed that agents were aware of their location $\boldsymbol{p}_k$. Let $d_{kq} = r_{kq} + \boldsymbol{x}_{kq}\boldsymbol{p}_k$, that is, $d_{kq} = \boldsymbol{x}_{kq}\boldsymbol{w}_q^o$. The problem was to estimate $\boldsymbol{w}_q^o$ from noisy streaming measurements $\{d_{kq}(n), \boldsymbol{x}_{kq,n}\}$ collected by each agent k, and governed by the linear model [13]:

$$d_{kq}(n) = \boldsymbol{x}_{kq,n}\boldsymbol{w}_q^o + z_{kq}(n)$$

with

$$\boldsymbol{x}_{kq,n} = [1 - \beta_k(n)]\,\boldsymbol{x}_{kq} + \boldsymbol{x}_{kq}^{\perp}\,\mathrm{diag}\{\alpha_{k1}(n), \alpha_{k2}(n)\} \tag{124}$$

with $z_{kq}(n)$ a zero-mean temporally and spatially i.i.d. Gaussian noise of variance σ_z^2. As shown in (124), the measured direction vector $\boldsymbol{x}_{kq,n}$ was assumed to be a noisy realization of the unit-norm vector pointing from $\boldsymbol{p}_k$ to $\boldsymbol{w}_q^o$, with $\boldsymbol{x}_{kq}^{\perp}$ a unit-norm orthogonal contribution to $\boldsymbol{x}_{kq}$. Random variables $\alpha_{k1}(n)$, $\alpha_{k2}(n)$, $\beta_k(n)$ and $z_k(n)$ were zero-mean Gaussian with standard deviation $\sigma_{\alpha_1} = \sigma_{\alpha_2} = 0.1$, $\sigma_\beta = 0.001$ and $\sigma_z = 0.3$, respectively. We ran the (non-cooperative) LMS algorithm at each node, and Algorithm 1, with $\boldsymbol{\Theta} = \boldsymbol{R}_{1,2}$ and $\boldsymbol{\Theta}_\perp = \boldsymbol{r}_3$. The step-size μ was set to 0.1. A uniform combination matrix $\boldsymbol{A}$ with $a_{\ell k} = |\mathcal{N}_k|^{-1}$ was used for Algorithm 1, where $|\mathcal{N}_k|$ denotes the cardinality of the neighborhood of node k. Figure 8(a) compares the MSD of these strategies. Figures 8(b) and 8(c) show one realization of the target locations estimated with the (noncooperative) LMS algorithm and Algorithm 1. This experiment illustrates the advantage of cooperative strategies over the non-cooperative one.

VI. Conclusion and Perspectives

In this paper, we formulated an online multitask adaptation problem that assumes that all tasks share a common latent feature representation, locally refined by node-specific contributions. This model can be extended into interesting directions by imposing new constraints, depending on applications. Based on this principle, we derived two cooperative algorithms and analyzed their performance. Although this work considers that common representation subspaces are known a priori, it paves the way towards more general frameworks.

Appendix A
Proof of Lemma 1

The uniqueness of the solution of (8) follows from the strict convexity of (10), which is ensured by the positive definiteness of its Hessian matrix. For the quadratic cost (10), the Hessian matrix with respect to the vector of stacked variables $\mathrm{col}\{\boldsymbol{u}, \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_N\}$ is block diagonal [3, App. B], with blocks given by the following matrix $\boldsymbol{X}$ and its transpose:

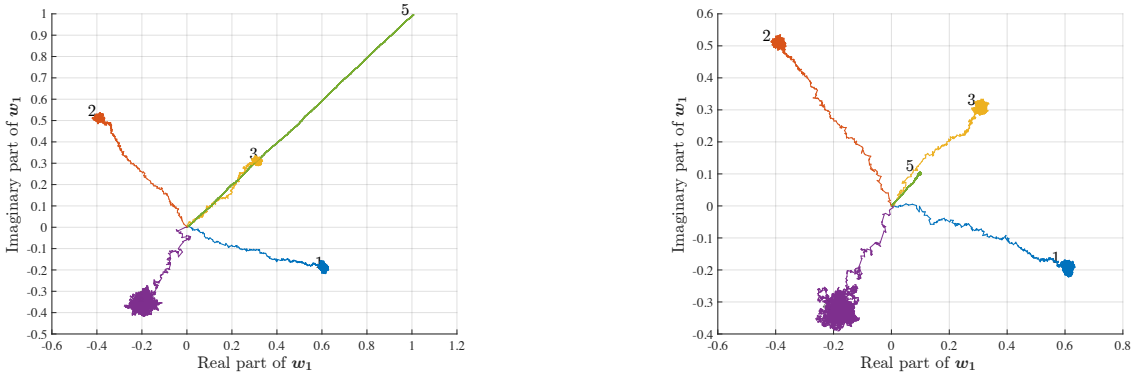
$$\nabla^2 J^{\mathrm{glob}} = \begin{bmatrix} \boldsymbol{X} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{X}^\top \end{bmatrix}$$

with

$$\boldsymbol{X} = \left(\begin{array}{c|ccc} \boldsymbol{\Theta}^*\big(\sum_{k=1}^N \boldsymbol{R}_{x,k}\big)\boldsymbol{\Theta} & \boldsymbol{\Theta}^*\boldsymbol{R}_{x,1}\boldsymbol{\Theta}_\perp & \cdots & \boldsymbol{\Theta}^*\boldsymbol{R}_{x,N}\boldsymbol{\Theta}_\perp \\ \hline \boldsymbol{\Theta}_\perp^*\boldsymbol{R}_{x,1}\boldsymbol{\Theta} & \boldsymbol{\Theta}_\perp^*\boldsymbol{R}_{x,1}\boldsymbol{\Theta}_\perp & & \boldsymbol{0} \\ \vdots & & \ddots & \\ \boldsymbol{\Theta}_\perp^*\boldsymbol{R}_{x,N}\boldsymbol{\Theta} & \boldsymbol{0} & & \boldsymbol{\Theta}_\perp^*\boldsymbol{R}_{x,N}\boldsymbol{\Theta}_\perp \end{array}\right) \tag{125}$$

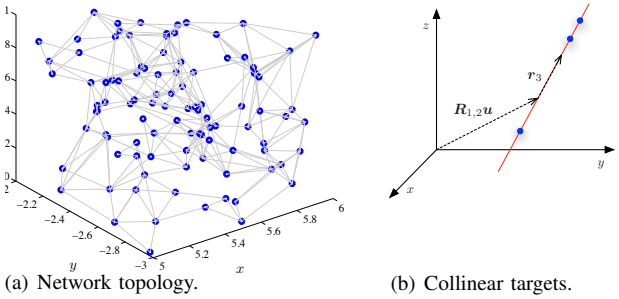
where $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}_\perp$ have full column rank. The positive definiteness of (125) can be checked by verifying the positive definiteness of each term $\boldsymbol{\Theta}^*\boldsymbol{R}_{x,k}\boldsymbol{\Theta}$ and of the Schur complement

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2017.2671789, IEEE Journal of Selected Topics in Signal Processing



(a) Weight behavior of Algorithm 1. (b) Weight behavior of Algorithm 2.

Fig. 6. Weight behavior of Algorithms 1 and 2 with singular inputs and non-zero mean disturbance.



(a) Network topology. (b) Collinear targets.

Fig. 7. Network topology and locations of targets.

relative to the block diagonal corner of $\boldsymbol{X}$, namely, [53]

$$\text{Schur}(\boldsymbol{X}) = \sum_{k=1}^{N} \big[\boldsymbol{\Theta}^*\boldsymbol{R}_{x,k}\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\boldsymbol{R}_{x,k}\boldsymbol{\Theta}_\perp(\boldsymbol{\Theta}_\perp^*\boldsymbol{R}_{x,k}\boldsymbol{\Theta}_\perp)^{-1}\boldsymbol{\Theta}_\perp^*\boldsymbol{R}_{x,k}\boldsymbol{\Theta}\big] \tag{126}$$

where each inverse $(\boldsymbol{\Theta}_\perp^*\boldsymbol{R}_{x,k}\boldsymbol{\Theta}_\perp)^{-1}$ exists since $\boldsymbol{\Theta}_\perp$ has full column rank. Each term inside the summation (126) is positive definite since it is the Schur complement of the block $\boldsymbol{\Theta}^*\boldsymbol{R}_{x,k}\,\boldsymbol{\Theta}$ in the positive definite matrix:

$$\begin{pmatrix} \boldsymbol{\Theta}^*\boldsymbol{R}_{x,k}\boldsymbol{\Theta} & \boldsymbol{\Theta}^*\boldsymbol{R}_{x,k}\boldsymbol{\Theta}_\perp \\ \boldsymbol{\Theta}_\perp^*\boldsymbol{R}_{x,k}\boldsymbol{\Theta} & \boldsymbol{\Theta}_\perp^*\boldsymbol{R}_{x,k}\boldsymbol{\Theta}_\perp \end{pmatrix} = [\boldsymbol{\Theta}\ \ \boldsymbol{\Theta}_\perp]^*\boldsymbol{R}_{x,k}[\boldsymbol{\Theta}\ \ \boldsymbol{\Theta}_\perp] > 0 \tag{127}$$

This guarantees the positive definiteness of (126). It follows that the cost in (10) is strictly convex and has a unique minimizer.

Appendix B
Proof of Lemma 2

Without loss of generality, assume that $\eta_1 > \eta_2$. Otherwise, replace (129) by:

$$J^{\text{glob}}\big(\boldsymbol{u}, \{\boldsymbol{\epsilon}_k\}_{k=1}^N\big) = \sum_{k=1}^{N} \mathbb{E}\big\{|d_k(n) - \boldsymbol{x}_{k,n}(\boldsymbol{\Theta}\boldsymbol{u} + \boldsymbol{\epsilon}_k)|^2\big\} + \eta_1 \sum_{k=1}^{N} \|\boldsymbol{\epsilon}_k\|^2 + (\eta_2 - \eta_1) \sum_{k=1}^{N} \|\boldsymbol{P}_{\boldsymbol{\Theta}_\perp}\boldsymbol{\epsilon}_k\|^2 \tag{128}$$

Recalling that $\boldsymbol{P}_{\boldsymbol{\Theta}_\perp} = \boldsymbol{I}_L - \boldsymbol{P}_{\boldsymbol{\Theta}}$, the objective function (31) can be written as follows:

$$J^{\text{glob}}\big(\boldsymbol{u}, \{\boldsymbol{\epsilon}_k\}_{k=1}^N\big) = (\eta_1 - \eta_2) \sum_{k=1}^{N} \|\boldsymbol{P}_{\boldsymbol{\Theta}}\boldsymbol{\epsilon}_k\|^2 + \underbrace{\sum_{k=1}^{N} \mathbb{E}\big\{|d_k(n) - \boldsymbol{x}_{k,n}(\boldsymbol{\Theta}\boldsymbol{u} + \boldsymbol{\epsilon}_k)|^2\big\} + \eta_2 \sum_{k=1}^{N} \|\boldsymbol{\epsilon}_k\|^2}_{J_1^{\text{glob}}} \tag{129}$$

The uniqueness of the minimizer of (31) follows from its strict convexity. For the quadratic cost in (129), the Hessian of J_1^{glob} with respect to the vector of stacked variables $\text{col}\{\boldsymbol{u}, \boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_N\}$ is again block diagonal, with its blocks determined by the matrix $\boldsymbol{Y}$ below and its transpose:

$$\nabla^2 J_1^{\text{glob}} = \begin{bmatrix} \boldsymbol{Y} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Y}^\top \end{bmatrix} \tag{130}$$

with

$$\boldsymbol{Y} = \left(\begin{array}{c|ccc} \boldsymbol{\Theta}^*\big(\sum_{k=1}^N \boldsymbol{R}_{x,k}\big)\boldsymbol{\Theta} & \boldsymbol{\Theta}^*\boldsymbol{R}_{x,1} & \cdots & \boldsymbol{\Theta}^*\boldsymbol{R}_{x,N} \\ \hline \boldsymbol{R}_{x,1}\boldsymbol{\Theta} & \boldsymbol{R}_{x,1} + \eta_2\,\boldsymbol{I} & & \boldsymbol{0} \\ \vdots & & \ddots & \\ \boldsymbol{R}_{x,N}\boldsymbol{\Theta} & \boldsymbol{0} & & \boldsymbol{R}_{x,N} + \eta_2\,\boldsymbol{I} \end{array}\right) \tag{131}$$

The positive definiteness of (130) can be checked by verifying the positive definiteness of each term $\boldsymbol{R}_{x,k} + \eta_2\,\boldsymbol{I}$ and of the Schur complement relative to the right block diagonal corner in (130), namely, [53]

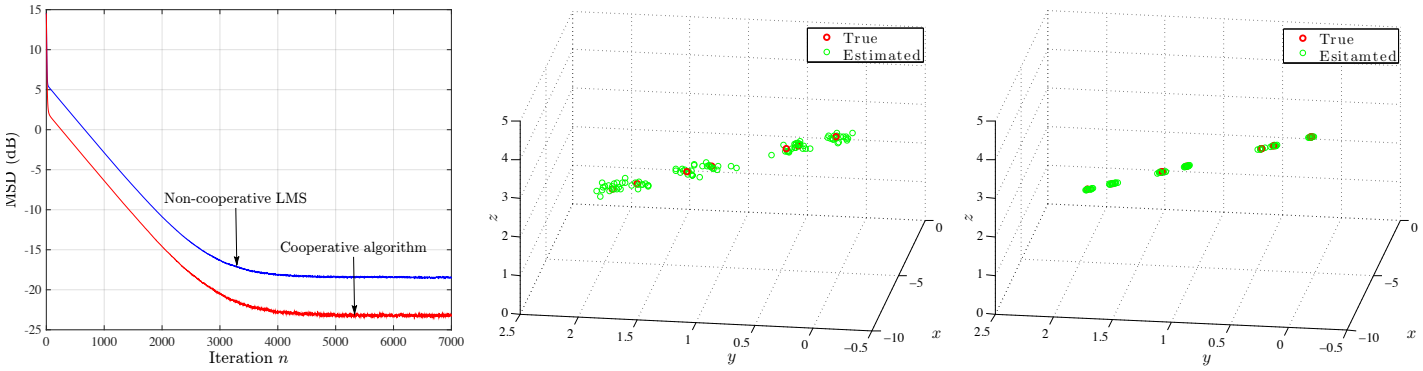
$$\text{Schur}(\boldsymbol{Y}) = \sum_{k=1}^{N} \big[\boldsymbol{\Theta}^*\boldsymbol{R}_{x,k}\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\boldsymbol{R}_{x,k}(\boldsymbol{R}_{x,k} + \eta_2\boldsymbol{I})^{-1}\boldsymbol{R}_{x,k}\boldsymbol{\Theta}\big] \tag{132}$$

Since they are positive definite, each covariance matrix $\boldsymbol{R}_{x,k}$ can be decomposed as follows:

$$\boldsymbol{R}_{x,k} = \boldsymbol{U}_k\,\text{diag}\{\lambda_{k,1}, \ldots, \lambda_{k,L}\}\,\boldsymbol{U}_k^* \tag{133}$$

where the $\lambda_{k,i}$ are the eigenvalues of $\boldsymbol{R}_{x,k}$, which are real and positive, and $\boldsymbol{U}_k$ is the corresponding matrix of eigenvectors. Since $\boldsymbol{U}_k$ is an orthonormal matrix, each term in the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JSTSP.2017.2671789, IEEE Journal of Selected Topics in Signal Processing



(a) Network MSD. (b) Estimation results (non-coop. LMS). (c) Estimation results (Algorithm 1).

Fig. 8. Estimated network MSD and estimation results for a single realization.

summation (132) can be written as:

$$\begin{aligned}&\boldsymbol{\Theta}^*\boldsymbol{R}_{x,k}\boldsymbol{\Theta}-\boldsymbol{\Theta}^*\boldsymbol{R}_{x,k}(\boldsymbol{R}_{x,k}+\eta_2\boldsymbol{I})^{-1}\boldsymbol{R}_{x,k}\boldsymbol{\Theta}\\ =&\boldsymbol{\Theta}^*\boldsymbol{U}\mathrm{diag}\left\{\lambda_{k,1}-\frac{\lambda_{k,1}^2}{\lambda_{k,1}+\eta_2},\ldots,\lambda_{k,L}-\frac{\lambda_{k,L}^2}{\lambda_{k,L}+\eta_2}\right\}\boldsymbol{U}^*\boldsymbol{\Theta}>0\end{aligned} \tag{134}$$

Since $\boldsymbol{\Theta}$ has full column rank, the above matrix and the Schur complement (132) are positive definite. In addition, the block diagonal matrix $\mathrm{diag}\{\boldsymbol{R}_{x,1}+\eta_2\boldsymbol{I},\ldots,\boldsymbol{R}_{x,N}+\eta_2\boldsymbol{I}\}$ is positive definite. Finally, since $(\eta_1-\eta_2)\sum_{k=1}^N\|\boldsymbol{P}_{\boldsymbol{\Theta}}\boldsymbol{\epsilon}_k\|^2$ in (129) is convex, problem (129) is strictly convex and problem (30) has a unique solution.

REFERENCES

[1] J. Chen, C. Richard, A. O. Hero, and A. H. Sayed, "Diffusion LMS for multitask problems with overlapping hypothesis subspaces," in *Proc. IEEE Int. Workshop on Machine Learn. for Signal Process. (MLSP)*, Reims, France, Sept. 2014, pp. 1–6.

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

[3] A. H. Sayed, "Adaptation, learning, and optimization over networks," in *Foundations and Trends in Machine Learning*, vol. 7, pp. 311–801. NOW Publishers, Boston-Delft, Jul. 2014.

[4] 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.

[5] D. P. Bertsekas, "A new class of incremental gradient methods for least squares problems," *SIAM J. Optimiz.*, vol. 7, no. 4, pp. 913–926, Nov. 1997.

[6] M. G. Rabbat and R. D. Nowak, "Quantized incremental algorithms for distributed optimization," *IEEE J. of Sel. Topics Areas Commun.*, vol. 23, no. 4, pp. 798–808, Apr. 2005.

[7] D. Blatt, A. O. Hero, and H. Gauchman, "A convergent incremental gradient method with constant step size," *SIAM J. Optimiz.*, vol. 18, no. 1, pp. 29–51, Feb. 2007.

[8] 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.

[9] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Trans. Autom. Control*, vol. 54, no. 1, pp. 48–61, Jan. 2009.

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

[11] K. Srivastava and A. Nedic, "Distributed asynchronous constrained stochastic optimization," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 4, pp. 772–790, Aug. 2011.

[12] A. H. Sayed, S.-Y Tu, J. Chen, X. Zhao, and Z. Towfic, "Diffusion strategies for adaptation and learning over networks," *IEEE Sig. Process. Mag.*, vol. 30, no. 3, pp. 155–171, May 2013.

[13] A. H. Sayed, "Diffusion adaptation over networks," in *Academic Press Library in Signal Processing*, R. Chellapa and S. Theodoridis, Eds., pp. 323–454. Academic Press, Elsevier, 2014.

[14] C. G. Lopes and A. H. Sayed, "Diffusion least-mean squares over adaptive networks: Formulation and performance analysis," *IEEE Trans. Signal Process.*, vol. 56, no. 7, pp. 3122–3136, Jul. 2008.

[15] 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.

[16] J. Chen and A. H. Sayed, "Diffusion adaptation strategies for distributed optimization and learning over networks," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 4289–4305, Aug. 2012.

[17] J. Chen and A. H. Sayed, "Distributed Pareto optimization via diffusion strategies," *IEEE J. Sel. Topics Signal Process.*, vol. 7, no. 2, pp. 205–220, Apr. 2013.

[18] 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.

[19] J. Chen, C. Richard, and A. H. Sayed, "Diffusion LMS over multitask networks," *IEEE Trans. Signal Process.*, vol. 63, no. 11, pp. 2733–2748, Jun. 2015.

[20] R. Abdolee, B. Champagne, and A. H. Sayed, "Estimation of space-time varying parameters using a diffusion LMS algorithm," *IEEE Trans. Signal Process.*, vol. 62, no. 2, pp. 403–418, Jan. 2014.

[21] Y. Hirata, D. P. Mandic, H. Suzuki, and K. Aihara, "Wind direction modelling using multiple observation points," *Philosophical Transactions of the Royal Society A*, vol. 366, pp. 591–607, 2008.

[22] R. Nassif, C. Richard, A. Ferrari, and A. H. Sayed, "Multitask diffusion LMS with sparsity-based regularization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Brisbane, Australia, 2015, pp. 3516–3520.

[23] X. Zhao and A. H. Sayed, "Distributed clustering and learning over networks," *IEEE Trans. Signal Process.*, vol. 63, no. 13, pp. 3285–3300, Jul. 2015.

[24] J. Chen, C. Richard, and A. H. Sayed, "Adaptive clustering for multitask diffusion networks," in *Proc. European Signal Process. Conf. (EUSIPCO)*, Nice, France, Sept. 2015, pp. 200–204.

[25] R. Nassif, C. Richard, J. Chen, A. Ferrari, and A. H. Sayed, "Diffusion LMS over multitask networks with noisy links," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Shanghai, China, 2016.

[26] S. Monajemi, K. Eftaxias, S. Sanei, and Ong S.-H., "An informed multitask diffusion adaptation approach to study tremor in Parkinson's disease," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 7, pp. 1306–1314, Oct. 2016.

[27] Y. Wang, W. P. Tay, and W. Hu, "Multitask diffusion LMS with optimized inter-cluster cooperation," in *Proc. IEEE Stat. Signal Process. Workshop (SSP)*, Palma de Mallorca, Spain, 2016, pp. 1–5.

[28] N. Roula, 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.

[29] S. Monajemi, S. Sanei, Ong S.-H., and A. H. Sayed, "Adaptive regularized diffusion adaptation over multitask networks," in *Proc. IEEE Int. Workshop on Machine Learn. for Signal Process. (MLSP)*, Boston, USA, Sept. 2015, pp. 1–5.

[30] S. Khawatmi, A. M. Zoubir, and A. H. Sayed, "Decentralized clustering over adaptive networks," in *Proc. European Signal Process. Conf. (EUSIPCO)*, Nice, France, 2015, pp. 2696–2700.

[31] J. Plata-Chaves, N. Bogdanović, 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.

[32] J. Plata-Chaves, H. H. Bahari, M. Moonen, and A. Bertrand, "Unsupervised diffusion-based LMS for node-specific parameter estimatin over wireless sensor networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 4159–4163.

[33] A. Bertrand and M. Moonen, "Distributed adaptive node-specific signal estimation in fully connected sensor networks – Part I: sequential node updating," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5277–5291, Oct. 2010.

[34] A. Bertrand and M. Moonen, "Distributed adaptive estimation of node-specific signals in wireless sensor networks with a tree topology," *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 2196–2210, May 2011.

[35] N. Bogdanović, J. Plata-Chaves, and K. Berberidis, "Distributed incremental-based LMS for node-specific adaptive parameter estimation," *IEEE Trans. Signal Process.*, vol. 62, no. 20, pp. 5382–5397, Oct. 2014.

[36] J. Chen, S. K. Ting, C. Richard, and A. H. Sayed, "Group diffusion LMS," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Shanghai, China, March 2016.

[37] J. Baxter, "A model of inductive bias learning," *J. Artif. Intell. Res.*, vol. 12, no. 1, pp. 149–198, Feb. 2000.

[38] R. K. Ando and T. Zhang, "A framework for learning predictive structures from multiple tasks and unlabeled data," *J. Mach. Learn. Res.*, vol. 6, pp. 1817–1853, 2005.

[39] J. Zhou, J. Chen, and J. Ye, "Clustered multi-task learning via alternating structure optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, Granada, Spain, Dec. 2011, pp. 702–710.

[40] J. Chen, L. Tang, J. Liu, and J. Ye, "A convex formulation for learning shared structures from muliple tasks," in *Proc. Ann. Int. Conf. Machine Learning (ICML)*, Montreal, Canada, Jun. 2009, pp. 137–144.

[41] L. Jacob, F. Bach, and J.-P. Vert, "Clustered multitask learning: A convex formulation," *Adv. Neural Inf. Process. Syst.*, vol. 21, pp. 745–752, 2009.

[42] G. Obozinski, B. Taskar, and M. Jordan, "Joint covariate selection and joint subspace selection for multiple classification problems," *Stat. Comput.*, vol. 20, no. 2, pp. 231–252, 2010.

[43] L. J. Griffiths and C. W. Jim, "An alternative approach to linear constrained adaptive beamforming," *IEEE Trans. Antenn. Propag.*, vol. 30, no. 1, pp. 27–34, Jan. 1982.

[44] X. Zhao and A. H. Sayed, "Clustering via diffusion adaptation over networks," in *Proc. Int. Workshop Cognitive Inf. Process. (CIP)*, Parador de Baiona, Spain, May 2012, pp. 1–6.

[45] R. Nassif, C. Richard, A. Ferrari, and A. H. Sayed, "Performance analysis of multitask diffusion adaptation over asynchronous networks," in *Proc. Asilomar Conf. Signals, Systems and Computers (ASILOMAR)*, Pacific Grove, CA, USA, Nov. 2014, pp. 788–792.

[46] J. Chen and C. Richard, "Performance analysis of diffusion LMS in multitask networks," in *Proc. IEEE Int. Workshop on Compt. Adv. in Multi-Sensor Adaptive Process. (CAMSAP)*, Saint Martin, France, Dec. 2013, pp. 137–140.

[47] J. Plata-Chaves, N. Bogdanović, and K. Berberidis, "Distributed incremental-based RLS for node-specific parameter estimation over adaptive networks," in *Proc. European Signal Process. Conf. (EUSIPCO)*, Marrakech, Morocco, Sept. 2013, pp. 1–5.

[48] N. Bogdanović, J. Plata-Chaves, and K. Berberidis, "Distributed diffusion-based LMS for node-specific parameter estimation over adaptive networks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Florence, Italy, May 2014, pp. 7223–7227.

[49] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[50] I. Gohberg and V. Olshevsky, "Fast algorithms with preprocessing for matrix-vector multiplication problems," *J. Complexity*, vol. 10, no. 4, pp. 411–427, 1994.

[51] A. H. Sayed, *Fundamentals of Adaptive Filtering*, J. Wiley & Sons, Hoboken, NJ, 2003.

[52] V. H. Nascimento and A. H. Sayed, "Unbiased and stable leakage-based adaptive filters," *IEEE Trans. Signal Process.*, vol. 47, no. 12, pp. 3261–3276, 1999.

[53] F. Zhang, *The Schur Complement and its Applications*, Springer, 2005.

Jie Chen (S'12–M'14) is a Professor at Center of Intelligent Acoustics and Immersive Communications (CIAIC) of Northwestern Polytechnical University. He received the Dipl.-Ing. and the M.S. degrees in 2009 from Troyes University of Technology (UTT), France, and from Xi'an Jiaotong University, respectively, all in information and telecommunication engineering. In 2013, he received the Ph.D. degree from UTT. From 2013 to 2014, he was with the Lagrange Laboratory, University of Nice Sophia Antipolis, France. From 2014 to 2015, he was with the Department of Electrical Engineering and Computer Science at the University of Michigan, Ann Arbor, USA.

Jie Chen has been recognized with the "Thousand Talents Plan (Youth Program)" Award in China. He was the Technical Co-Chair of IWAENC'16 held in Xi'an, China. His current research interests include adaptive signal processing, distributed optimization with applications in hyperspectral image analysis, acoustic signal processing and bioinformatics.

Cédric Richard (S'98–M'01–SM'07) received the Dipl.-Ing. and the M.S. degrees in 1994, and the Ph.D. degree in 1998, from Compiègne University of Technology, France. He is a Full Professor at the University of Nice Sophia Antipolis, France. He was a junior member of the Institut Universitaire de France in 2010-2015.

His current research interests include statistical signal processing and machine learning. Prof. Richard is the author of over 250 papers. He was the General Co-Chair of the IEEE SSP'11 Workshop that was held in Nice, France. He was the Technical Co-Chair of EUSIPCO'15 that was held in Nice, France, and of the IEEE CAMSAP'15 Workshop that was held in Cancun, Mexico. Since 2015, he serves as a Senior Area Editor of the IEEE Transactions on Signal Processing, and as an Associate Editor of the IEEE Transactions on Signal and Information Processing over Networks. He is an Associate Editor of Signal Processing Elsevier since 2009. Prof. Richard is member of the IEEE Machine Learning for Signal Processing Technical Committee, and served as member of the IEEE Signal Processing Theory and Methods Technical Committee in 2009-2014.

Ali H. Sayed (S'90–M'92–SM'99–F'01) is distinguished professor and former chairman of electrical engineering at the University of California, Los Angeles, where he directs the UCLA Adaptive Systems Laboratory. An author of over 490 scholarly publications and six books, his research involves several areas including adaptation and learning, statistical signal processing, distributed processing, network science, and biologically-inspired designs. His work has been recognized with several awards including 2014 Athanasios Papoulis Award from the European Association for Signal Processing, the 2015 Education Award, the 2013 Meritorious Service Award, and the 2012 Technical Achievement Award from the IEEE Signal Processing Society, the 2005 Terman Award from the American Society for Engineering Education, the 2003 Kuwait Prize, and the 1996 IEEE Donald G. Fink Prize. He served as Distinguished Lecturer for the IEEE Signal Processing Society in 2005 and as Editor-in Chief of the IEEE Transactions on Signal Processing (2003-2005). His articles received several Best Paper Awards from the IEEE Signal Processing Society in 2002, 2005, 2012, 2014, 2015. He is a Fellow of both the IEEE and the American Association for the Advancement of Science (AAAS). He is recognized as a Highly Cited Researcher by Thomson Reuters.