Sparse Distributed Estimation via Heterogeneous Diffusion Adaptive Networks

Bijit K. Das, *Student Member, IEEE*, Mrityunjoy Chakraborty, *Senior Member, IEEE*, and Jerónimo Arenas-García, *Senior Member, IEEE*

***Abstract*—Recently, diffusion networks have been proposed to identify sparse linear systems which employ sparsity-aware algorithms like the zero-attracting LMS (ZA-LMS) at *each* node to exploit sparsity. In this brief, we show that the same optimum performance as reached by the aforementioned networks can also be achieved by a "heterogeneous" network with *only a fraction* of the nodes deploying ZA-LMS-based adaptation, provided that the ZA-LMS-based nodes are distributed over the network maintaining some "uniformity." Reduction in the number of sparsity-aware nodes reduces the overall computational burden of the network. We show analytically and also by simulation studies that the only adjustment needed to achieve this reduction is a proportional increase in the value of the optimum zero attracting coefficient.**

***Index Terms*—Adaptive network, diffusion LMS, network mean square deviation (NMSD), l_1 norm.**

I. Introduction

In RECENT years, diffusion strategies [1]–[4] have been widely used to carry out distributed estimation in real time. Here, each node employs an adaptive filter that updates a tap weight vector using some local input and desired response sequences and separately refines it before or after each adaptation cycle by using tap weight information from neighboring nodes. Separately, the topic of sparse adaptive filtering has assumed special significance in recent years, and several new techniques have been proposed. A review of this can be found in [5]. Recently, in [6] and [7], the diffusion principle has been used to identify sparse finite impulse response (FIR) systems, where certain sparsity-promoting norms like the l_1 norm of the coefficient vector have been used to regularize the standard LMS cost function. This leads to the deployment of the sparsity-aware zero-attracting LMS (ZA-LMS) ([8], [9]) form of weight adaptation at each node. These diffusion sparse LMS algorithms manifest superior performance in terms of lesser steady-state network mean square deviation (NMSD) compared with the simple diffusion LMS.

Manuscript received September 29, 2015; revised February 20, 2016; accepted March 23, 2016. Date of publication March 29, 2016; date of current version October 27, 2016. This brief was recommended by Associate Editor T. Ogunfunmi.

B. K. Das and M. Chakraborty are with the Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology, Kharagpur 721 302, India (e-mail: bijitbijit@gmail.com; mrityun@ece.iitkgp.ernet.in).

J. Arenas-García is with the Department of Signal Theory and Communications, Universidad Carlos III de Madrid, 28903 Madrid, Spain (e-mail: jarenas@tsc.uc3m.es).

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

Digital Object Identifier 10.1109/TCSII.2016.2548182

In this brief, we show that the minimum level of the steady-state NMSD achieved using a ZA-LMS based update at *all* the nodes of the network can also be obtained by a *heterogeneous* network with only a fraction of the nodes using the ZA-LMS update rule (referred to as sparsity-aware nodes in this brief) while rest employ the standard LMS update (referred to as sparsity agnostic nodes in this brief), provided that the sparsity-aware nodes are distributed over the network maintaining some "uniformity." Note that reduction in the number of sparsity-aware nodes reduces the overall computational overhead of the network, especially when more complicated sparsity-aware algorithms involving a significant amount of computations are deployed to exploit sparsity. We show that, to achieve this reduction, the only adjustment to be made is a proportional increase in the value of the optimum zero attracting coefficient. Lastly, the proposed analysis, though restricted to the l_1-norm regularized algorithm (i.e., ZA-LMS) only, can be trivially extended to the cases of more general norms, and thus, similar behavior can also be expected from the corresponding heterogeneous networks.[1]

II. Brief Review of Diffusion Sparse LMS Algorithms

We consider a connected network consisting of N nodes that are spatially distributed. At every time index n, each kth node collects some scalar measurement $d_k(n)$ and some $M \times 1$ vector $\mathbf{u}_k(n)$ which are related by the following model: $d_k(n) = \mathbf{u}_k^T(n)\mathbf{w}_0 + v_k(n)$, where $v_k(n)$ is the measurement noise at the kth node and $\mathbf{w}_0$ is the unknown $M \times 1$ vector, known *a priori* to be sparse, which is required to be estimated (jointly by all the nodes). Both $\mathbf{u}_k(n)$ and $v_k(n)$ are variates generated from some Gaussian distributions, with $\mathbf{u}_k(n)$ and $v_k(m)$ being mutually independent for all $n,\ m$.

In the diffusion scheme, every kth node, $k = 1, 2, \cdots, N$, deploys an $M \times 1$ adaptive filter $\mathbf{w}_k(n)$ to estimate $\mathbf{w}_0$, which takes $d_k(n)$ and $\mathbf{u}_k(n)$, respectively, as the local desired response and input vectors. The estimates of $\mathbf{w}_0$, i.e., $\mathbf{w}_k(n)$, for each k are exchanged with the neighbors of the kth node, i.e., nodes directly connected to it, and the estimates are refined by following either the "adapt-then-combine (ATC)" or the "combine-then-adapt (CTA)" [1] schemes. In this brief, we consider the ATC approach where $\mathbf{w}_k(n)$ is first updated to an intermediate estimate $\mathbf{v}_k(n+1)$, which is then linearly

[1] Some early results of this brief were presented at International Symposium on Circuits and Systems (ISCAS)-2015 [10].

TABLE I
ZA-ATC DIFFUSION ALGORITHM [7]

$$e_k(n) = d_k(n) - \mathbf{w}_k^T(n)\mathbf{u}_k(n)$$
$$\mathbf{v}_k(n+1) = \mathbf{w}_k(n) + \mu_k \mathbf{u}_k(n) e_k(n) - \rho_k sgn[\mathbf{w}_k(n)] \quad (1)$$
$$\mathbf{w}_k(n+1) = \sum_{j \in \aleph_k} g_{j,k} \mathbf{v}_j(n+1) \quad (2)$$

combined with similar estimates received from the neighbors. Also, for sparsity-aware nodes, we assume the ZA-LMS form of weight adaptation ([8], [9]), obtained by adding the l_1-norm penalty $\|\mathbf{w}_k(n)\|_1$ to the LMS cost function which results in the introduction of the zero attracting terms $\mathrm{sgn}[\mathbf{w}_k(n)]$ in the weight update equations. The resulting diffusion ZA-LMS algorithm for the ATC scheme, popularly termed as ZA-ATC diffusion algorithm [7], is shown in Table I, where ρ_k is a very small positive zero-attracting coefficient (taken the same for all the nodes in [7]) and $\aleph_k$ denotes the kth neighborhood, i.e., the set of nodes directly connected to node k (including itself), having a total of $N_k = |\aleph_k|$ nodes ($|\cdot|$ denotes the cardinality of the set "$\cdot$" in the argument).

The combining coefficients $g_{j,k}$ are nonnegative constants which are usually chosen satisfying the following [1]: $g_{j,k} > 0$ if $j \in \aleph_k$ and equals zero otherwise. Also, $\sum_{j\in\aleph_k} g_{j,k} = 1$. There exist several standard schemes in the literature to choose the coefficients $g_{j,k}$, e.g., the uniform combination rule, the metropolis rule, the Laplacian rule, and the nearest neighbor rule to name a few. Using these coefficients, a combination matrix $\mathbf{G}$ is defined for the network, where $[\mathbf{G}]_{j,k} = g_{j,k}$.

III. PROPOSED HETEROGENEOUS NETWORK AND ITS NMSD BEHAVIOR

The performance of a diffusion network is assessed by evaluating the average NMSD which, at the nth time index, is given as [1]–[3] $\mathrm{MSD}_{\mathrm{net}}(n) = (1/N)\sum_{k=1}^N \mathrm{MSD}_k(n)$, where $\mathrm{MSD}_k(n)$ is the individual mean square deviation at the kth node, i.e., $\mathrm{MSD}_k(n) = E[\|\tilde{\mathbf{w}}_k(n)\|^2]$ where $\tilde{\mathbf{w}}_k(n) = \mathbf{w}_0 - \mathbf{w}_k(n)$ denotes the weight deviation vector at the kth node. An expression for the steady-state NMSD [i.e., $\mathrm{MSD}_{\mathrm{net}}(\infty)$] for a general ZA-ATC algorithm was derived analytically in [7]. Under the simplifying assumptions that all nodes employ the same step size μ and that both the input signal and noise at each node are spatially and temporally independent and identically distributed (i.i.d.), it is easy to verify that the $\mathrm{MSD}_{\mathrm{net}}(\infty)$ in [7] reduces to the following:

$$\mathrm{MSD}_{\mathrm{net}}(\infty) = \frac{\mu^2\sigma_v^2\sigma_u^2}{N}\left[\mathrm{vec}[\mathbf{C}^T\mathbf{C}]\right]^T(\mathbf{I}-\mathbf{F})^{-1}\mathbf{q} + \frac{1}{N}\left(\beta(\infty) - \alpha(\infty)\right) \quad (3)$$

with

$$\alpha(\infty) = -2\,\mu E\left[\mathrm{sgn}\left[\mathbf{w}(\infty)\right]^T \boldsymbol{\Omega}_s \mathbf{C}\mathbf{C}^T(\mathbf{I}-\mu\mathbf{D})\tilde{\mathbf{w}}(\infty)\right] \quad (4)$$

$$\beta(\infty) = \mu^2 E\left[\|\mathrm{sgn}\left[\mathbf{w}(\infty)\right]\|^2_{\boldsymbol{\Omega}_s\mathbf{C}\mathbf{C}^T\boldsymbol{\Omega}_s}\right] \quad (5)$$

where $\mathrm{vec}[\cdot]$ is an operator that stacks the columns of its argument matrix on top of each other, $\mathbf{q} = \mathrm{vec}[\mathbf{I}_{MN\times MN}]$, $\mathbf{w}(n) = \mathrm{vec}[\mathbf{w}_1(n),\ \mathbf{w}_2(n), \cdots \mathbf{w}_N(n)]$, $\tilde{\mathbf{w}}(n) = \mathrm{vec}[\tilde{\mathbf{w}}_1(n), \tilde{\mathbf{w}}_2(n), \ldots \tilde{\mathbf{w}}_N(n)]$, and σ_v^2 and σ_u^2 are the variances of the noise and input signal, respectively. The matrices $\mathbf{C}$, $\mathbf{D}$, $\mathbf{F}$, and $\boldsymbol{\Omega}_s$ are defined as follows: $\mathbf{C} = \mathbf{G}\otimes\mathbf{I}_{M\times M}$ ($\otimes$ denotes the right Kronecker product), $\mathbf{D} = \sigma_u^2\mathbf{I}_{MN\times MN}$, $\mathbf{F} = (1 - 2\mu\sigma_u^2 + \mu^2\sigma_u^4)(\mathbf{C}\otimes\mathbf{C})$, and $\boldsymbol{\Omega}_s = \mathrm{diag}[\rho_1\mathbf{I}_{M\times M}, \rho_2\mathbf{I}_{M\times M}, \cdots \rho_k\mathbf{I}_{M\times M} \ldots \rho_N\mathbf{I}_{M\times M}]$. (Also note that, for a vector $\mathbf{a}$ and a matrix $\mathbf{B}$, $\|\mathbf{a}\|^2_{\mathbf{B}}$ indicates $\mathbf{a}^T\mathbf{B}\mathbf{a}$.)

It is easily seen that the first term on the right hand side (RHS) of (3) is actually the steady-state network mean square deviation (NMSD) of simple ATC diffusion LMS [3] and is independent of ρ. Now, in the case of the ZA-ATC diffusion algorithm [7], all nodes were assumed to employ the same zero attracting coefficient, i.e., it was assumed that $\rho_1 = \rho_2 = \cdots = \rho_N = \rho$ (say). In such case, the second term on the RHS of (3) can be written as $(1/N)\phi(\rho)$, where $\phi(\rho) = -\alpha'(\infty)\rho + \beta'(\infty)\rho^2$, with $\alpha'(\infty) = -2\mu E[\mathrm{sgn}[\mathbf{w}(\infty)]^T \mathbf{C}\mathbf{C}^T(\mathbf{I}-\mu\mathbf{D})\tilde{\mathbf{w}}(\infty)]$ and $\beta'(\infty) = \mu^2 E[\|\mathrm{sgn}[\mathbf{w}(\infty)]\|^2_{\mathbf{C}\mathbf{C}^T}] (> 0)$. The function $\phi(\rho)$ has two zero-crossing points, one at $\rho = 0$ and the other at $\rho = (\alpha'(\infty))/(\beta'(\infty))$, and between them, $\phi(\rho)$ takes only negative values with the minima occurring at $\rho = (\alpha'(\infty))/(2\beta'(\infty))$, which, from (3), also minimizes $\mathrm{MSD}_{\mathrm{net}}(\infty)$. For systems that are highly sparse, it follows from [7] that $\alpha'(\infty) > 0$, and conversely, for nonsparse systems, $\alpha'(\infty) < 0$. Since, for proper zero attraction, ρ must be positive, the optimum value of ρ is then given by $\rho_{\mathrm{opt}} = \max[0, (\alpha'(\infty))/(2\beta'(\infty))]$, and the corresponding minimum value of ϕ (when $\rho_{\mathrm{opt}} > 0$) is then given as

$$\phi_{\min} = -\frac{\alpha'^2(\infty)}{4N\beta'(\infty)}. \quad (6)$$

A. Proposed Heterogeneous Diffusion Network

In this section, we show that the same level of $\phi_{\min}$ as given by (6) and, therefore, the same $\min[\mathrm{MSD}_{\mathrm{net}}(\infty)]$ can be reached by a heterogeneous network as well, where only a fraction of the nodes are sparsity aware and the rest are sparsity agnostic, with the network satisfying the following assumptions as closely as possible (where we use the symbols S and $L_s (= |S|)$ to denote respectively the set of the sparsity-aware nodes and the number of sparsity-aware nodes in the network).

Assumption 1: The matrix $\mathbf{G}$ is symmetric and also doubly stochastic, i.e., $\forall i, j$, $\sum_{i=1}^N g_{i,j} = 1$ and $\sum_{j=1}^N g_{i,j} = 1$. This is valid for many practical rules used to select combiner coefficients [1].

Assumption 2: The L_s sparsity-aware nodes are distributed over the network uniformly such that, for any jth node ($1 \le j \le N$), the number of sparsity-aware nodes in its neighborhood is directly proportional to N_j. We take the proportionality constant in this case to be L_s/N, which is consistent with the above, since it implies that, if, hypothetically, $N_j = 1\ \forall j$ (i.e., there is a total of N neighborhoods with each neighborhood consisting of one node only), then each neighborhood should have a quota of L_s/N sparsity-aware nodes.

A network with strong connectivity is expected to satisfy Assumption 2 closely, as is also confirmed by our simulation studies. A fallout of this assumption is that, under the uniform combination rule [1] for which $g_{i,j} = (1/N_j)\ \forall i \in \aleph_j$, we will have $\sum_{i\in S} g_{i,j} = L_s/N$.[2] This is validated by our simulation studies in Section IV where we consider three networks of increasing order of connectivity and show that the above gets satisfied more closely as the connectivity of the network increases.

Now, in order to prove our conjecture that the same level of $\phi_{\min}$ as given by (6) can also be achieved in a heterogeneous network, we have to minimize $(1/N)(\beta(\infty) - \alpha(\infty))$ w.r.t. ρ. However, unlike [7], for the heterogeneous network, it is lot more difficult to express $\alpha(\infty)$ and $\beta(\infty)$ as a function of ρ since, unlike in [7], $\boldsymbol{\Omega}_s$ cannot be written simply as $\rho\mathbf{I}$. In Theorems 1 and 2, we evaluate $\alpha(\infty)$ and $\beta(\infty)$, respectively, in terms of ρ, where we make use of the following definitions.

Definition 1: Here, we define the matrix $\boldsymbol{\theta}$ by taking a weighted average of the matrix $E[\text{sgn}[\mathbf{w}_i(\infty)]\tilde{\mathbf{w}}_m^T(\infty)]$ over all $i \in S$ and for each i, considering every mth node that is connected with the ith node either directly (i.e., $m \in \aleph_i$) or through an intermediate jth node (i.e., $m \in \aleph_j, j \in \aleph_i,\ j \neq i$). In particular, we define

$$\boldsymbol{\theta} = \frac{\sum_{i\in S}\sum_{m\in\aleph_j, j\in\aleph_i} E\left[\text{sgn}\left[\mathbf{w}_i(\infty)\right]\tilde{\mathbf{w}}_m^T(\infty)\right]\mathbf{g}_i^T\mathbf{g}_m}{\sum_{i\in S}\sum_{m\in\aleph_j, j\in\aleph_i}\mathbf{g}_i^T\mathbf{g}_m}$$

where $\mathbf{g}_l$ is the lth column of the matrix $\mathbf{G}$ and the constants $\mathbf{g}_i^T\mathbf{g}_m$ are the weights used in the weighted average.

Definition 2: Similarly, we define the matrix $\boldsymbol{\psi}$ by taking a weighted average of $E[\text{sgn}[\mathbf{w}_i(\infty)]\text{sgn}[\mathbf{w}_m^T(\infty)]]$ over all $i \in S$ and for each i, considering every mth *sparsity-aware* node that is connected with the ith node either directly (i.e., $m \in S \cap \aleph_i$) or through an intermediate jth node (i.e., $m \in S \cap \aleph_j,\ j \in \aleph_i$, $j \neq i$). In particular, we define

$$\boldsymbol{\psi} = \frac{\sum_{i\in S}\sum_{m\in S\cap\aleph_j, j\in\aleph_i} E\left[\text{sgn}[\mathbf{w}_i(\infty)]\text{sgn}\left[\mathbf{w}_m^T(\infty)\right]\right]\mathbf{g}_i^T\mathbf{g}_m}{\sum_{i\in S}\sum_{m\in S\cap\aleph_j,\ j\in\aleph_i}\mathbf{g}_i^T\mathbf{g}_m}.$$

It is then possible to prove the following.

Theorem 1: For a network satisfying Assumption 1, the following holds:

$$\alpha(\infty) = -2\rho\mu\left(1 - \mu\sigma_u^2\right)Tr[\boldsymbol{\theta}]L_s. \quad (7)$$

Proof: Using the fact that for any two vectors $\mathbf{a}$ and $\mathbf{b}$, $\mathbf{a}^T\mathbf{b} \equiv \text{Tr}[\mathbf{a}\mathbf{b}^T]$ (Tr[.] denotes the trace of the argument matrix), one can write from (4), $\alpha(\infty) = -2\,\mu(1-\mu\sigma_u^2)\text{Tr}[\boldsymbol{\Omega}_s E[\text{sgn}[\mathbf{w}(\infty)]\tilde{\mathbf{w}}^T(\infty)]\mathbf{C}\mathbf{C}^T]$. Defining the $MN \times MN$ block matrices $\mathbf{K1} = E[\text{sgn}[\mathbf{w}(\infty)]\tilde{\mathbf{w}}^T(\infty)]$, $\mathbf{K2} = \mathbf{C}\mathbf{C}^T$, and $\mathbf{K} = \mathbf{K1}\mathbf{K2}$ and also denoting by $\mathbf{A}_{p,q}$ the (p,q)th subblock of an $MN \times MN$ block matrix $\mathbf{A}$, one can write $\alpha(\infty) = -2\,\mu(1-\mu\sigma_u^2)\text{Tr}[\boldsymbol{\Omega}_s\mathbf{K}] = -2\rho\mu(1-\mu\sigma_u^2)Tr[\sum_{i\in S}\mathbf{K}_{i,i}] = -2\rho\mu(1-\mu\sigma_u^2)Tr[\sum_{i\in S}\sum_{m=1}^N \mathbf{K1}_{i,m}\mathbf{K2}_{m,i}]$. Now, $\mathbf{K2}_{m,i} = \sum_{r=1}^N \mathbf{C}_{m,r}\mathbf{C}_{i,r}^T$. From the definition of $\mathbf{C}$, one can easily see that $\mathbf{C}_{i,r}^T = \mathbf{C}_{i,r} = g_{i,r}\mathbf{I}_N$. This leads to

$$\alpha(\infty) = -2\rho\mu\left(1-\mu\sigma_u^2\right)Tr\left[\sum_{i\in S}\times\sum_{m=1}^{N}E\left[\text{sgn}\left[\mathbf{w}_i(\infty)\right]\tilde{\mathbf{w}}_m^T(\infty)\right]\mathbf{g}_i^T\mathbf{g}_m\right]. \quad (8)$$

Now, if the mth node is directly connected to the ith node, meaning that $[\mathbf{g}_i]_m > 0$, or has, for example, the lth node as a common neighbor, meaning that $[\mathbf{g}_i]_l > 0$, $[\mathbf{g}_m]_l > 0$, the product $\mathbf{g}_i^T\mathbf{g}_m$ will be nonzero and positive, while in all other cases, it will be zero. In other words, $\mathbf{g}_i^T\mathbf{g}_m \neq 0$ if and only if $m \in \aleph_j,\ j \in \aleph_i$. Using this in (8) and then invoking the aforesaid Definition 1, one obtains $\alpha(\infty) = -2\rho\mu(1-\mu\sigma_u^2)Tr[\boldsymbol{\theta}]\sum_{i\in S}\sum_{m\in\aleph_j, j\in\aleph_i}\mathbf{g}_i^T\mathbf{g}_m$. Rewriting the summation $\sum_{i\in S}\sum_{m\in\aleph_j, j\in\aleph_i}\mathbf{g}_i^T\mathbf{g}_m$ as $\sum_{i\in S}\mathbf{g}_i^T[\sum_{m=1}^N \mathbf{g}_m]$ (i.e., using the orthogonality of $\mathbf{g}_m$ with $\mathbf{g}_i$ for m not belonging to $\aleph_j, j \in \aleph_i$), noting from Assumption 1 that $\sum_{m=1}^N \mathbf{g}_m = [1, 1, \cdots, 1]^T$ and also that $\mathbf{g}_i^T[1, 1, \cdots, 1]^T = 1$, one obtains the result as given in (7). □

Corollary 1: For the proposed network, $\text{Tr}[\boldsymbol{\theta}] < 0$ for highly sparse systems, and conversely, $\text{Tr}[\boldsymbol{\theta}] > 0$ for nonsparse systems.

Proof: In [7], it was shown that, for networks using the same zero attracting coefficient ρ at all the nodes, i.e., $\rho_i = \rho,\ i = 1, 2, \cdots, N$, the following holds: $\alpha(\infty) > 0$ for highly sparse systems, and for nonsparse systems, $\alpha(\infty) < 0$. For the generalized case under consideration, i.e., when $\rho_1 \neq \rho_2 \neq \cdots \neq \rho_N$ with each $\rho_i \geq 0$, one obtains $\alpha(\infty)$ [as given by the RHS of (4)] simply by replacing $\rho\text{sgn}[\mathbf{w}(\infty)]$ in the corresponding derivation in [7] by $\boldsymbol{\Omega}_s\text{sgn}[\mathbf{w}(\infty)]$. Since each $\rho_i \geq 0$ with at least one $\rho_i > 0$, from [7], it then follows that, for the generalized case also, we have $\alpha(\infty) > 0$ for highly sparse systems and $\alpha(\infty) < 0$ for nonsparse systems. From this and the fact that, for convergence, we have $1 - \mu\sigma_u^2 > 0$, the result then follows trivially from Theorem 1. □

In a similar way, we evaluate $\beta(\infty)$ as given in the theorem hereinafter.

Theorem 2: For a network satisfying the aforesaid Assumptions 1 and 2, the following holds:

$$\beta(\infty) = \frac{\mu^2\rho^2 Tr[\boldsymbol{\psi}]L_s^2}{N}. \quad (9)$$

Proof: As in Theorem 1, we first define the $MN \times MN$ block matrices $\mathbf{L1} = E[\text{sgn}[\mathbf{w}(\infty)]\text{sgn}[\mathbf{w}(\infty)]^T]$, $\mathbf{L2} = \mathbf{L1}\boldsymbol{\Omega}_s$, and $\mathbf{L} = \mathbf{L2}\mathbf{K2}$, where $\mathbf{K2} = \mathbf{C}\mathbf{C}^T$, i.e., same as defined in the proof of Theorem 1. From (5), one can write $\beta(\infty) = \mu^2 E[\text{sgn}[\mathbf{w}(\infty)]^T\boldsymbol{\Omega}_s\mathbf{C}\mathbf{C}^T\boldsymbol{\Omega}_s\,\text{sgn}[\mathbf{w}(\infty)]] \equiv \mu^2\,\text{Tr}[\boldsymbol{\Omega}_s E[\text{sgn}[\mathbf{w}(\infty)]\,\text{sgn}[\mathbf{w}(\infty)]^T]\boldsymbol{\Omega}_s\mathbf{C}\mathbf{C}^T]$. Using the

[2]The aforementioned relation is satisfied (albeit approximately) for another widely popular combination rule, namely, the metropolis rule [1]. Under this, for $i \in \aleph_j$ but $i \neq j$, $g_{i,j} = 1/(\max[|\aleph_i|, |\aleph_j|])$, and for $i = j$, $g_{j,j} = 1 - \sum_{i\in\aleph_j, i\neq j} g_{i,j}$ (note that, under this rule, $g_{i,j} = g_{j,i}$, i.e., the symmetry of $\mathbf{G}$ is automatically satisfied). Since, for a strongly connected network, the number of nodes in a neighborhood does not vary much from neighborhood to neighborhood, i.e., $\aleph_k \approx \aleph_l$ for any $k, l \in \{1, 2, \cdots, N\}$, we will have $g_{i,j} \approx 1/N_j$, and thus, $\sum_{i\in S} g_{i,j} \approx L_s/N$.

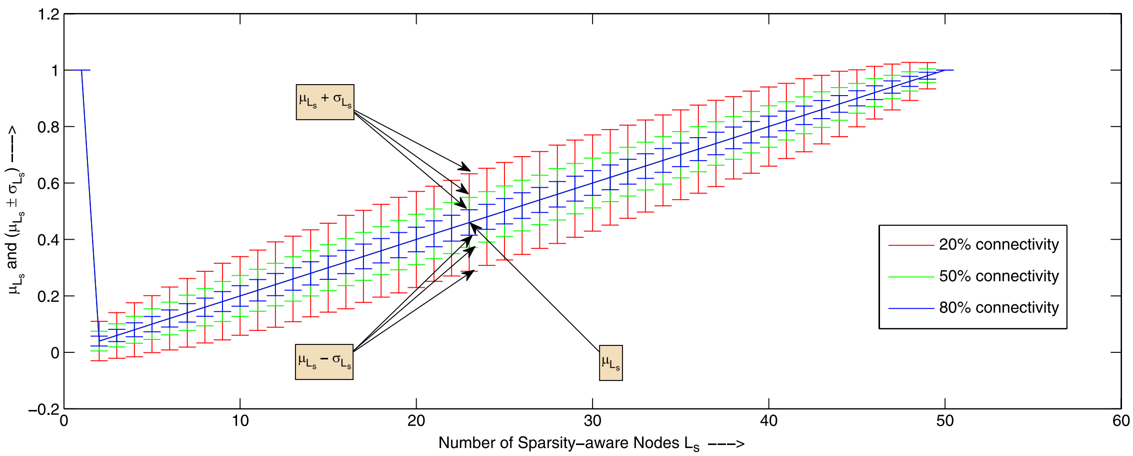

Fig. 1. Variation of μ_{L_s} and $\mu_{L_s} \pm \sigma_{L_s}$ against the number of sparsity-aware nodes (L_s).

aforementioned definitions of block matrices and following the steps used in the proof of Theorem 1, one can then express $\beta(\infty)$ as $\beta(\infty) = \mu^2 \mathrm{Tr}[\boldsymbol{\Omega}_s \mathbf{L}] = \mu^2 \rho \mathrm{Tr}[\sum_{i \in S} \mathbf{L}_{i,i}] = \mu^2 \rho Tr\left[\sum_{i \in S} \sum_{m=1}^{N} \mathbf{L2}_{i,m} \mathbf{K2}_{m,i}\right] = \mu^2 \rho^2 \mathrm{Tr}\left[\sum_{i \in S} \sum_{m \in S} \mathbf{L1}_{i,m} \mathbf{K2}_{m,i}\right]$. Recalling from Theorem 1 that $\mathbf{K2}_{m,i} = \sum_{r=1}^{N} \mathbf{C}_{m,r} \mathbf{C}_{i,r}^T$ and that $\mathbf{C}_{i,r}^T = \mathbf{C}_{i,r} = g_{i,r} \mathbf{I}_N$, $\mathbf{C}_{m,r} = g_{m,r} \mathbf{I}_N$, one obtains

$$\beta(\infty) = \mu^2 \rho^2 Tr\left[\sum_{i \in S} \times \sum_{m \in S} E\left[\mathrm{sgn}\left[\mathbf{w}_i(\infty)\right] \mathrm{sgn}\left[\mathbf{w}_m^T(\infty)\right]\right] \mathbf{g}_i^T \mathbf{g}_m\right]. \quad (10)$$

Using the fact that $\mathbf{g}_i^T \mathbf{g}_m \neq 0$ if and only if $m \in \aleph_j$, $j \in \aleph_i$, and from the aforesaid Definition 2, one obtains $\beta(\infty) = \mu^2 \rho^2 Tr[\boldsymbol{\psi}] \sum_{i \in S} \sum_{m \in S \cap \aleph_j, j \in \aleph_i} \mathbf{g}_i^T \mathbf{g}_m$. Next, we rewrite the summation $\sum_{i \in S} \sum_{m \in S \cap \aleph_j, j \in \aleph_i} \mathbf{g}_i^T \mathbf{g}_m$ as $\sum_{i \in S} \mathbf{g}_i^T [\sum_{m \in S} \mathbf{g}_m]$. Then, for any k, $1 \leq k \leq N$, we have $[\sum_{m \in S} \mathbf{g}_m]_k = \sum_{m \in S} g_{k,m} = \sum_{m \in S} g_{m,k}$ (from the symmetric nature of $\mathbf{G}$) $= L_s/N$ (from Assumption 2). Thus, $\sum_{m \in S} \mathbf{g}_m = (L_s/N)[1, 1, \cdots, 1]^T$, and from Assumption 1, $\sum_{i \in S} \mathbf{g}_i^T [\sum_{m \in S} \mathbf{g}_m] = L_s^2/N$. By substitution, one then obtains the result as given in (9). □

Substituting $\alpha(\infty)$ and $\beta(\infty)$ in $\phi(\rho) = (1/N)(\beta(\infty) - \alpha(\infty))$, then differentiating w.r.t. ρ, and equating the derivative to zero, we obtain

$$\rho_{\mathrm{opt}} = \max\left[0, -\frac{\left(1 - \mu \sigma_u^2\right) \mathrm{Tr}[\boldsymbol{\theta}] N}{\mu \mathrm{Tr}[\boldsymbol{\psi}] L_s}\right]. \quad (11)$$

The corresponding minimum value of $\phi(\rho)$ (when $\rho_{\mathrm{opt}} > 0$, i.e., the system is sparse), say, $\phi'_{\min}$, is given as

$$\phi'_{\min} = -\frac{\left(1 - \mu \sigma_u^2\right)^2 \mathrm{Tr}[\boldsymbol{\theta}]^2}{\mathrm{Tr}[\boldsymbol{\psi}]}. \quad (12)$$

Note that $\phi'_{\min}$ as given in (12) is independent of L_s. Several conclusions follow from this as given below.

- Since $\phi'_{\min}$ as given in (12) is true for all values of L_s, it holds also for $L_s = N$, i.e., when the network becomes *homogeneous with all the nodes being sparsity aware*. In other words, if $\phi_{\min}$ as given by (6) is analyzed using the aforesaid Assumptions 1 and 2, it would give rise to the same expression as given by (12).
- The $\min[\mathrm{MSD}_{\mathrm{net}}(\infty)]$ does not change when the network changes from being homogeneous to heterogeneous, with only L_s of the total N $(0 < L_s < N)$ nodes employing sparsity-aware adaptation.
- For sparse systems, the ρ_{opt} minimizing $\phi(\rho)$ and, thus, $[\mathrm{MSD}_{\mathrm{net}}(\infty)]$ as given by (12) is, however, inversely proportional to L_s.

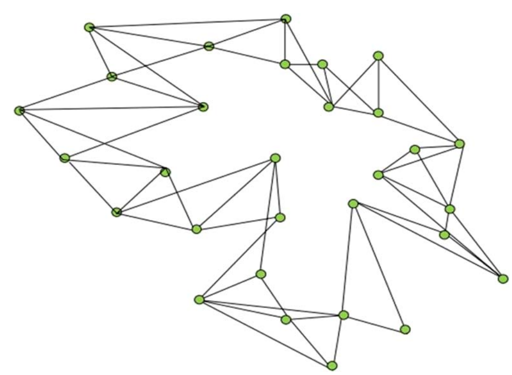

Fig. 2. Topology of the network used for simulation (i.e., Fig. 3).

The above implies that the same $\phi'_{\min}$ as realized by the homogeneous network can also be reached by the heterogeneous network for various $\{\rho_{\mathrm{opt}}, L_s\}$ combinations, with ρ_{opt} being inversely proportional to L_s. In other words, it is possible to bring down the computational complexity of the network while maintaining the same $\phi'_{\min}$ by choosing a lesser value of L_s with a proportionate increase in ρ.

IV. Simulation Studies

In our simulation exercise, we first try to validate Assumption 2 made in the previous section. In particular, we try to show that, under well-known combination rules like the uniform combination or metropolis rule, we have $\sum_{i \in S} g_{i,j} = L_s/N$. For this, we consider a network of $N = 50$ nodes and form a 50×50 connectivity matrix, first with 20% connectivity,

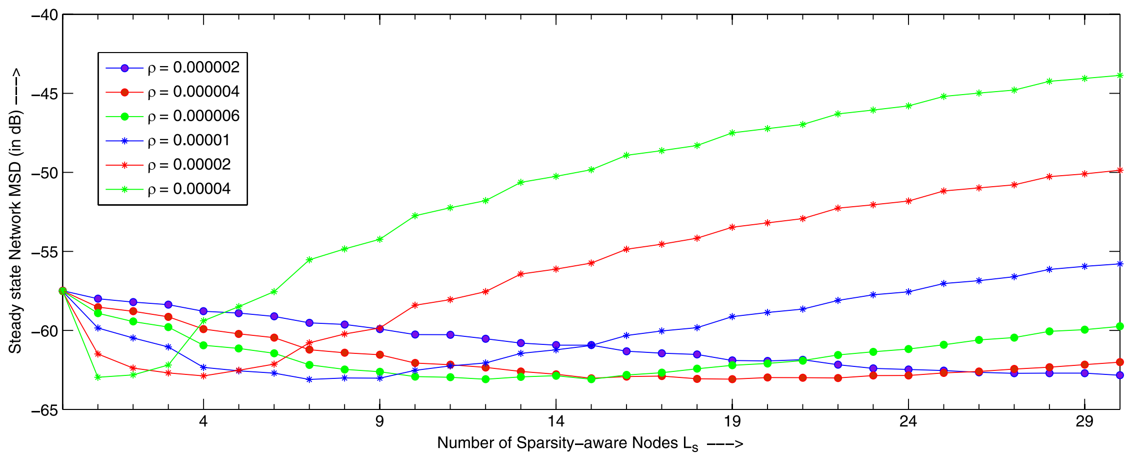

Fig. 3. Network MSD versus number of sparsity-aware nodes (L_s) for different values of ρ.

i.e., having only 20% of the links between all possible pairs of nodes in place. Then, using the metropolis rule, the combination weights are determined. Next, we select a subset S of the given N nodes randomly to be the set of sparsity-aware nodes, and for the chosen S, calculate $\sum_{i\in S} g_{i,j}$ for each jth neighborhood $(1 \leq j \leq N)$ and evaluate their spatial mean and variance. Then, keeping the size $|S|$ $(= L_s)$ fixed, the experiment is run 100 times, and an ensemble average of the aforementioned spatial mean and spatial variance, for example, μ_{L_s} and $\sigma^2_{L_s}$, respectively, are calculated. This is done for each value of L_s in the range of 1–50.

The aforementioned experiment is then repeated for 50% and 80% connectivities. The simulation results are displayed in Fig. 1 by plotting μ_{L_s} and the two levels, $\mu_{L_s} \pm \sigma_{L_s}$ (i.e., approximate range of fluctuation) vis-a-vis L_s, for each level of connectivities. Two observations follow easily from Fig. 1: i) μ_{L_s} for any L_s is given by L_s/N irrespective of the level of connectivities, and ii) as the connectivity increases, $\sigma^2_{L_s}$ decreases for any L_s. This validates Assumption 2.

Next, we consider a strongly connected network of $N = 30$ nodes with a topology as shown in Fig. 2. The weights of the network edges are assigned by the uniform combination rule [1]. Since all the nodes have more or less the same number of neighbors, the symmetry and doubly stochastic characteristics of the combination matrix $\mathbf{G}$ are easily satisfied. Also, nodes for sparsity-aware adaptation are chosen carefully so that the aforesaid Assumption 2 is satisfied as closely as possible. The same step size of $\mu = 6 \times 10^{-3}$ is chosen for all the nodes. Also, the input as well as measurement noise is taken from identical Gaussian distribution with $\sigma_u^2 = 1$, and $\sigma_v^2 = 1 \times 10^{-4}$, respectively across all the nodes. The goal of the network is to estimate a 128×1 vector $\mathbf{w}_0$ which is highly sparse (only one coefficient being nonzero). To estimate this, the value of ρ is first kept fixed at 2×10^{-6} for all the L_s sparsity-aware nodes, and L_s is varied over 0 to 30. For each L_s chosen, the adaptation of the network is carried out for 3000 iterations, and the steady-state NMSD is evaluated by taking the ensemble average over 1000 independent runs. The NMSD is then plotted as a function of L_s. Next, the value of ρ is increased progressively, taking the following values: 4×10^{-6}, 6×10^{-6}, 1×10^{-5}, 2×10^{-5} and 4×10^{-5}, and for each value of ρ chosen, the same exercise is carried out.

Fig. 3 displays the steady-state NMSD vis-a-vis L_s with ρ as a parameter. The following can be easily observed from Fig. 3: i) The minimum value reached by each NMSD-versus-L_s plot is the same for all the plots, and ii) as ρ increases, the value of L_s where the minimum occurs reduces and vice versa. In other words, Fig. 3 validates the theoretical conjectures made in this brief.

REFERENCES

[1] A. H. Sayed, "Diffusion adaptation over networks," in *E-Reference Signal Processing*, R. Chellapa and S. Theodoridis, Eds. Amsterdam, The Netherlands: Elsevier, 2013.

[2] 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.

[3] 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.

[4] S.-H. Yim, H.-S. Lee, and W.-J. Song, "A proportionate diffusion LMS algorithm for sparse distributed estimation," *IEEE Trans. Circuits Syst. II, Exp. Brief*, vol. 62, no. 10, pp. 992–996, Oct. 2015.

[5] R. L. Das and M. Chakraborty, "Sparse adaptive filters—An overview and some new results," in *Proc. ISCAS*, May 2012, Seoul, South Korea.

[6] Y. Liu, C. Li, and Z. Zhang, "Diffusion sparse least-mean squares over networks," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 4480–4485, Aug. 2012.

[7] P. Di Lorenzo and A. H. Sayed, "Sparse distributed learning based on diffusion adaptation," *IEEE Trans. Signal Process.*, vol. 61, no. 6, pp. 1419–1433, Mar. 15 2013.

[8] Y. Gu, Y. Chen, and A. O. Hero, "Sparse LMS for system identification," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Taipei, Taiwan, Apr. 2009, pp. 3125–3128,.

[9] K. Shi and P. Shi, "Convergence analysis of sparse LMS algorithms with l_1-norm penalty based on white input signal," *Signal Process.*, vol. 90, pp. 3289–3293, Dec. 2010.

[10] B. K. Das, M. Chakraborty, and J. Arenas-García, "Sparse distributed learning via heterogeneous diffusion adaptive networks," in *Proc. ISCAS*, May 2015, Lisbon, Portugal, pp. 437–440.