DIFFUSION LEAST-MEAN SQUARES OVER ADAPTIVE NETWORKS

Cassio G. Lopes and *Ali H. Sayed*

Department of Electrical Engineering
University of California
Los Angeles, CA, 90095.
Email: {cassio, sayed@ee.ucla.edu}

ABSTRACT

Distributed adaptive algorithms are proposed to address the problem of estimation in distributed networks. We extend recent work [1]–[2] by relying on static and adaptive diffusion strategies. The resulting adaptive networks are robust to node and link failures and present a substantial improvement over the non-cooperative case asserting that cooperation improves estimation performance. The distributed algorithms are peer-to-peer implementations suitable for networks with general topologies.

***Index Terms*—** Adaptive filters, Distributed estimation, Adaptive estimation, Adaptive signal processing, Cooperative systems

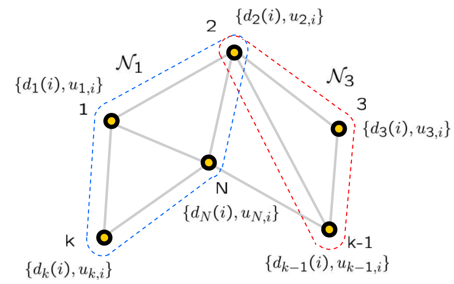

Fig. 1. A distributed network with N nodes.

1. INTRODUCTION

Distributed networks embedded with cooperative algorithms have been proposed to address estimation problems that arise in a variety of applications, such as environment monitoring, target localization and potential sensor network problems [4].

Recent work has been proposed to address distributed estimation problems adaptively by relying on limited cooperation and incremental-like techniques [1, 2]. If computational complexity is not an issue, more sophisticated learning rules may be employed such as distributed RLS algorithms [3]. When more communication resources are available, distributed adaptive algorithms can be derived that exploit more fully the network connectivity and increase the degree of cooperation among nodes. In this work we propose an adaptive LMS-like rule for a diffusion cooperation protocol, resulting in cooperative peer-to-peer estimation schemes that are suitable to operate over general topology networks.

2. DIFFUSION LMS

We want to estimate an $M \times 1$ unknown vector w^o from measurements collected at N nodes in a network (see Fig.1). Each node k has access to time realizations $\{d_k(i), u_{k,i}\}$, $k = 1, \ldots, N$, of zero-mean random data $\{\boldsymbol{d}_k, \boldsymbol{u}_k\}$, with $d_k(i)$ a scalar measurement and $u_{k,i}$ a regression row vector; both at time i. We resort to the concept of adaptive networks, recently investigated in [1, 2, 3], to perform the estimation task. An adaptive network results from equipping the nodes of the network with local learning rules, or local adaptive filters. The available communication topology is then employed to implement a cooperation protocol among the nodes in order to efficiently exploit spatial and temporal information. Different learning rules allied with different cooperation protocols give rise to different adaptive networks.

In this work we motivate and analyze a simple yet efficient aggregate-and-adapt diffusion protocol. Assume each node k has an unbiased estimate ψ_k^{i-1} of w^o at time $i-1$. This estimate can be interpreted as a noisy version of w^o, say $\psi_k^{i-1} = w^o - \tilde{\psi}_k^{i-1}$ for some error $\tilde{\psi}_k^{i-1}$. The neighborhood $\mathcal{N}_k$ of node k is defined as the set of all nodes linking to it, including itself. By linearly combining the estimates at the neighborhood of k we may replace ψ_k^{i-1} by a weighted estimate

$$\phi_k^{i-1} = \sum_{l \in \mathcal{N}_k} c_l \psi_l^{i-1}$$

for some combination coefficients $\{c_l \geq 0\}$. This aggregate estimate at node k can be interpreted as a weighted least-squares estimate of w^o given the $\{\psi_l^{i-1}\}$ at all neighbors of node k. The aggregation step helps fuse information from nodes across the network (and not just from the neighborhood $\mathcal{N}_k$) into node k. This is because generally every node in $\mathcal{N}_k$ tends to have a different neighborhood for connected topologies – see Fig. 1. The resulting aggregate ϕ_k^{i-1} at node k can subsequently be fed into a local adaptive filter in order to respond to local information and update it to ψ_k^i. Analysis and simulations will show that this scheme leads to a robust and fast distributed adaptive system that achieves smaller error levels in steady-state than its non-cooperative counterpart (where each node in the network adapts independently of other nodes and of aggregation).

The proposed diffusion strategy may be described in general

This material was based on work supported in part by the National Science Foundation under award ECS-0601266. The work of Mr. Lopes was also supported by a fellowship from CAPES, Brazil, under award 1168/01-0.

terms as follows:

$$\phi_k^{i-1} = f_k\left(\psi_\ell^{i-1}; \ell \in \mathcal{N}_{k,i-1}\right) \tag{1}$$

$$\psi_k^{i} = \phi_k^{i-1} + \mu_k u_{k,i}^*\left(d_k(i) - u_{k,i}\phi_k^{i-1}\right) \tag{2}$$

for some local combiner $f_k(\cdot)$ and step-size μ_k. The combiners may be nonlinear or even time variant, to reflect, for instance, changing topologies or to respond to non-stationary environments. The neighborhoods $\mathcal{N}_k$ may also be time-variant. The resulting adaptive network is a peer-to-peer estimation framework that is robust to node and link failures and exploits network connectivity.

In order to illustrate the technique, we explore a linear combiner model. At node k, the aggregated estimate ϕ_k^{i-1} is generated by linearly combining the neighbors' estimates, i.e.,

$$\begin{aligned} \phi_k^{i-1} &= \sum_{\ell \in \mathcal{N}_k} c(k,\ell)\, \psi_\ell^{i-1} \\ \psi_k^{i} &= \phi_k^{i-1} + \mu_k u_{k,i}^*\left(d_k(i) - u_{k,i}\phi_k^{i-1}\right) \end{aligned} \tag{3}$$

for a set of *local* combiners c_k satisfying $\sum_\ell c(k,\ell) = 1$ and assuming fixed neighborhoods $\mathcal{N}_k$.

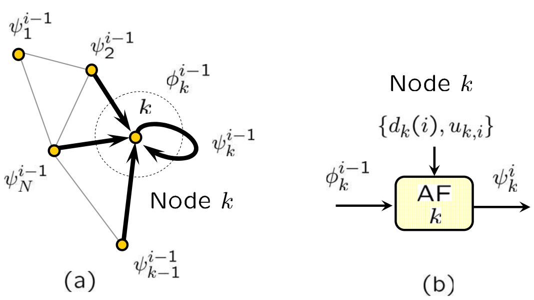

Fig. 2. A network with diffusion cooperation strategy.

3. STABILITY IN THE MEAN

Algorithm (3) embeds the combined effect of several adaptive filter updates, in addition to the network topology. Hence, performance analysis is challenging. However, the following analysis sheds some interesting insights on the role of cooperation and network topology on system performance. We resort to state-space representations. We introduce the global quantities:

$$\begin{aligned} &\boldsymbol{\psi}^i \triangleq \text{col}\{\boldsymbol{\psi}_1^i, \ldots, \boldsymbol{\psi}_N^i\}, \quad \boldsymbol{\phi}^{i-1} \triangleq \text{col}\{\boldsymbol{\phi}_1^{i-1}, \ldots, \boldsymbol{\phi}_N^{i-1}\} \\ &\boldsymbol{U}_i \triangleq \text{diag}\{\boldsymbol{u}_{1,i}, \ldots, \boldsymbol{u}_{N,i}\}, \quad \boldsymbol{d}_i = \text{col}\{\boldsymbol{d}_1(i), \ldots, \boldsymbol{d}_N(i)\} \end{aligned}$$

in terms of the stochastic quantities whose realizations appear in (3). Let $D = \text{diag}\{\mu_1 I_M, \mu_2 I_M, \ldots, \mu_N I_M\}$ be a diagonal matrix collecting the local step-sizes. The measurements are assumed to obey the linear model [5]:

$$\boldsymbol{d}_k(i) = \boldsymbol{u}_{k,i} w^o + \boldsymbol{v}_k(i) \tag{4}$$

where $\boldsymbol{v}_k(i)$ is background noise, assumed independent over time and space and with variance $\sigma_{v,k}^2$. A global representation for (3) is then given by

$$\begin{aligned} \boldsymbol{\phi}^{i-1} &= G\boldsymbol{\psi}^{i-1} \\ \boldsymbol{\psi}^{i} &= \boldsymbol{\phi}^{i-1} + D\boldsymbol{U}_i^*\big(\boldsymbol{d}_i - \boldsymbol{U}_i\boldsymbol{\phi}^{i-1}\big) \end{aligned} \tag{5}$$

or, in a more compact state-space form:

$$\boldsymbol{\psi}^{i} = G\boldsymbol{\psi}^{i-1} + D\boldsymbol{U}_i^*\big(\boldsymbol{d}_i - \boldsymbol{U}_i G\boldsymbol{\psi}^{i-1}\big) \tag{6}$$

where $G = C \otimes I_M$ is the transition matrix and $C = [c(k,\ell)]$ is a diffusion combination matrix. The matrix C has information about the network topology: a nonzero entry $c(k,\ell)$ means that node k is connected to node ℓ. Moreover, C satisfies $Cq_N = q_N$, where $q_N \triangleq \text{col}\{1, \ldots, 1\}$. Let

$$w^{(o)} \triangleq q_N \otimes w^o \quad \text{and} \quad \widetilde{\boldsymbol{\psi}}^i \triangleq w^{(o)} - \boldsymbol{\psi}^i \tag{7}$$

Furthermore, from the data model (4) we have $\boldsymbol{d}_i = \boldsymbol{U}_i w^{(o)} + \boldsymbol{v}_i$, with $\boldsymbol{v}_i = \text{col}\{\boldsymbol{v}_1(i), \ldots, \boldsymbol{v}_N(i)\}$. Moreover, since $Gw^{(o)} = w^{(o)}$, by subtracting $w^{(o)}$ from the left side and $Gw^{(o)}$ from the right side of (6) we get

$$\begin{aligned} \widetilde{\boldsymbol{\psi}}^i &= Gw^{(o)} - G\boldsymbol{\psi}^{i-1} - D\boldsymbol{U}_i^*\left(\boldsymbol{U}_i w^{(o)} + \boldsymbol{v}_i - \boldsymbol{U}_i G\boldsymbol{\psi}^{i-1}\right) \\ &= G\widetilde{\boldsymbol{\psi}}^{i-1} - D\boldsymbol{U}_i^*\left(\boldsymbol{U}_i G\widetilde{\boldsymbol{\psi}}^{i-1} + \boldsymbol{v}_i\right) \end{aligned} \tag{8}$$

or, equivalently,

$$\boxed{\widetilde{\boldsymbol{\psi}}^i = \left(I_{NM} - D\boldsymbol{U}_i^*\boldsymbol{U}_i\right) G\widetilde{\boldsymbol{\psi}}^{i-1} - D\boldsymbol{U}_i^*\boldsymbol{v}_i} \tag{9}$$

Assuming temporal and spatial independence of the regressors and taking expectations of both sides leads to

$$E\widetilde{\boldsymbol{\psi}}^i = \big(I_{NM} - DR_u\big)G\, E\widetilde{\boldsymbol{\psi}}^{i-1} \tag{10}$$

where $R_u = \text{diag}\{R_{u,1}, \ldots, R_{u,N}\}$ is block diagonal and $R_{u,k} = E\boldsymbol{u}_{k,i}^*\boldsymbol{u}_{k,i}$. Henceforth, for stability in the mean we must have that

$$\big|\lambda\big((I_{NM} - DR_u)G\big)\big| < 1 \tag{11}$$

In other words, the spectrum of $(I_{NM} - DR_u)G$ must be strictly inside the unit disc. In the absence of cooperation (i.e., when nodes evolve independently of each other), the mean error vector would evolve according to

$$E\widetilde{\boldsymbol{\psi}}^i = \big(I_{NM} - DR_u\big)\, E\widetilde{\boldsymbol{\psi}}^{i-1}$$

with coefficient matrix $B = (I_{NM} - DR_u)$. Thus, in the adaptive network case, even convergence in the mean will effectively depend on space-time data statistics *and* network topology (represented by G). For simplicity, assume that $D = \mu I_{NM}$ so that B is Hermitian. Using matrix 2-norms we have

$$\|BG\|_2 \;\le\; \|B\|_2 \cdot \|G\|_2 \tag{12}$$

That is, $\sigma_{\max}(BG) \le \sigma_{\max}(B) \cdot \sigma_{\max}(G)$ [6], where $\sigma_{\max}$ is the maximum singular value of the corresponding matrix. For combiners that render stochastic and symmetric matrices C, the matrix G will also be symmetric and stochastic so that $\sigma_{\max}(G) = 1$. But since $\sigma_{\max}(B) = |\lambda_{\max}(B)|$ and $|\lambda_{\max}(BG)| \le \sigma_{\max}(BG)$, we conclude that

$$\boxed{|\lambda_{\max}(BG)| \le |\lambda_{\max}(B)|} \tag{13}$$

That is, the spectral radius of BG is generally smaller than the spectral radius of B. Hence, cooperation under the diffusion protocol (3) has a *stabilizing* effect on the network.

One such combiner is the Metropolis rule. Let n_k and n_ℓ be the degree for nodes k and ℓ, i.e., $n_k = |\mathcal{N}_k|$. We have

$$\begin{cases} c(k,\ell) = 1/\max(n_k, n_\ell) \quad \text{if } k \neq \ell \text{ are } \textit{linked} \\ c(k,\ell) = 0 \quad \text{if } k \text{ and } \ell \text{ not } \textit{linked} \\ c(k,k) = 1 - \sum_{\ell \in \mathcal{N}_{k}/k} c(k,\ell) \quad \text{for } k = \ell \end{cases} \tag{14}$$

Other possible rules are the Laplacian and the nearest neighbor rules, given respectively by

$$C = I_N - \kappa L \qquad \text{and} \qquad c(k,\cdot) = \frac{1}{|\mathcal{N}_k|} \tag{15}$$

where $L = \mathcal{D} - A_d$, with $\mathcal{D} = \text{diag}\{n_1, \ldots, n_N\}$, A_d is the network Laplacian and $\kappa = 1/n_{\max}$. The Laplacian rule renders a contracting matrix [7], hence also decreasing the network convergence modes (See Fig. 7, right plot).

Naturally, convergence in the mean is only a necessary condition for convergence in the mean-square sense, which will be addressed in a complementary paper, but it gives important insights about the network behavior operating under diffusion protocols.

4. ADAPTIVE DIFFUSION LMS

This strategy can be understood as an adaptive layer implemented over the existing network of adaptive filters, constituting an *adaptive diffusion* protocol. Instead of combining "blindly" the estimates from the neighborhood, a better policy is to weigh them according to their respective individual performance.

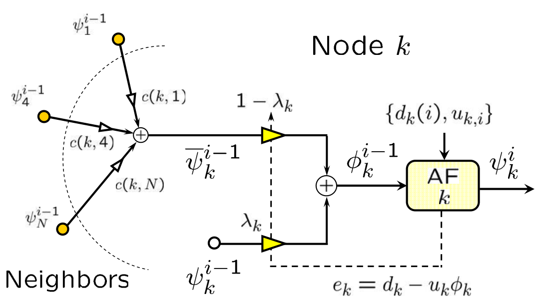

Fig. 3. A simple adaptive diffusion strategy.

There are many ways to design the combiner f_k in (1). We extend [8] to the network case. Each node k generates ϕ_k^{i-1} as

$$\phi_k^{i-1} \quad = \quad \lambda_k \psi_k^{i-1} + (1-\lambda_k)\,\overline{\psi}_k^{i-1} \tag{16}$$

where $\overline{\psi}_k^{i-1}$ is a linear combination of the neighbors' estimates

$$\overline{\psi}_k^{i-1} \quad = \quad \sum_{\ell \in \mathcal{N}_k/k} c(k,\ell)\psi_\ell^{i-1} \tag{17}$$

The resulting ϕ_k^{i-1} is presented to the local adaptive filter

$$e_k(i) \quad = \quad d_k(i) - u_{k,i}\phi_k^{i-1} \tag{18}$$

$$\psi_k^i \quad = \quad \phi_k^{i-1} + \mu\, u_{k,i}^* e_k(i) \tag{19}$$

The convex combiner $\lambda_k(a_k) \in [0,1]$ is a real activation function at our choice and depends on a parameter a_k that is adapted to minimize the local error (see Fig. 3), say as[1]

$$a_k \quad = \quad a_k - \mu_a \left[\nabla_a |e_k|^2\right]^* \tag{20}$$

[1]Time index will be dropped for compactness.

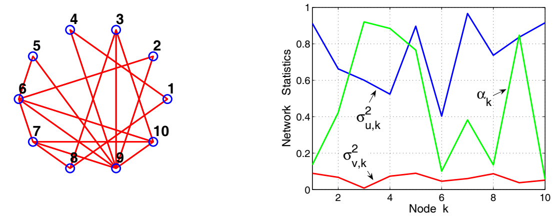

Fig. 4. Example 1: Network topology and statistical profile.

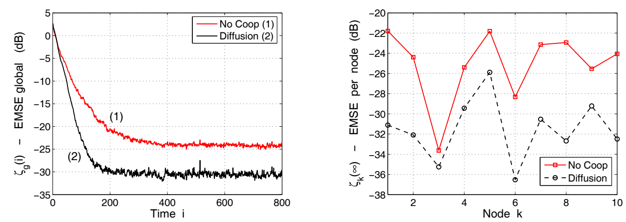

Fig. 5. Transient global EMSE and steady-state EMSE per node.

Choosing $\lambda_k = \frac{1}{1+\left|\exp(-a_k/2)\right|^2}$ and using (16)–(19) results in

$$a_k \quad = \quad a_k + \mu_a \frac{u_{k,i}^*}{\|u_{k,i}\|^\beta}\left(\psi_k^{i-1} - \overline{\psi}_k^{i-1}\right) e_k(i)\lambda_k\,(1-\lambda_k)$$

where β is a non-negative normalization parameter inspired by normalized LMS algorithms [5].

There are different ways to construct adaptive diffusion algorithms. We present here a pilot design to illustrate the concept: multilevel and clustered adaptation may substantially benefit the whole network.

We first compare diffusion LMS with the non-cooperative case, in which the adaptive filters evolve independently consulting only their own past estimates. We use the global average excess mean-square error (EMSE) as a figure of merit, defined as

$$\zeta_g(i) = \frac{1}{N}\sum_{k=1}^{N} \zeta_k(i)$$

where $\zeta_k(i) = E\,|\boldsymbol{u}_{k,i}(w^o - \boldsymbol{\psi}_k^{i-1})|^2$ is the local EMSE at node k. We also examine the network performance in steady-state by inspecting $\zeta_k(\infty)$ at every node. Network topology and statistics are randomly generated and 100 experiments are performed in each example. The regressors follow a local first order Markov process with power $\sigma_{u,k}^2$ and correlation index α_k. We adopt the Metropolis rule for the combiners.

For the first example we consider a network with $N = 10$ nodes whose topology and statistics are depicted in Fig. 4. Every node runs adaptive filters with $\mu = 0.015$ and $M = 10$ taps. Figure 5 (left) shows the global learning behavior and the network individual EMSE profile in steady-state (right). Note how the diffusion network presents a better performance.

In the second example we investigate the robustness of the diffusion protocol. A network with $N = 12$ nodes is set up with adaptive filters containing $M = 10$ taps, $\mu = 0.04$ and $\sigma_v^2 = 10^{-3}$. The topology and statistical profile are presented in Fig. 6. The step-size μ was not carefully designed, so that in the non-cooperative scheme some nodes did not converge uniformly, causing the global

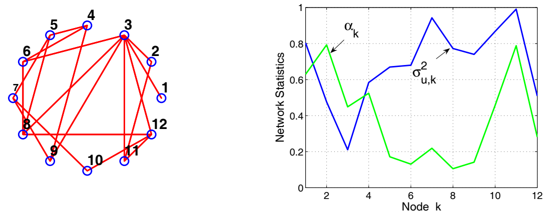

Fig. 6. Example 2: Network topology and statistical profile.

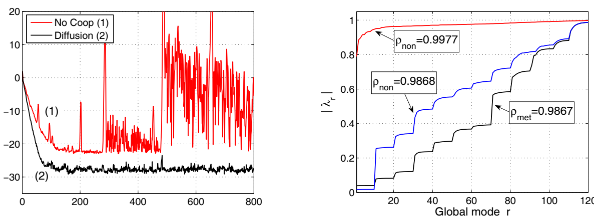

Fig. 7. Transient global EMSE and the spectrum of both non-cooperative and diffusion cases.

network error to increase. Nevertheless, in the diffusion cooperative scheme the global error remained under control and no spikes were observed. As Fig. 7 shows (left), cooperation is more robust and leads to improvement in network performance. The right plot in Fig. 7 reveals the spectrum (11) of both non-cooperative and diffusion protocols for the Laplacian and Metropolis rules. Note how the eigenmodes of the diffusion LMS are substantially smaller than the non-cooperative case. The Metropolis rule is even smaller than the Laplacian. The spectral radius ρ of both diffusion rules is smaller than the non-cooperative case: the diffusion protocol has a stabilizing effect over the network.

The operation of the adaptive diffusion scheme is illustrated in Figs. 8 and 9. For this example $\sigma_v^2 = 10^{-3}$, $\mu = 0.01$, $\mu_a = 40$ and $\beta = \{2, 6\}$. Note in Fig. 9 (left) that adaptive diffusion is faster than the standard diffusion protocol, but with slightly larger error (5dB @ -55dB) due to the extra adaptive layer (gradient noise). However this effect can be balanced by designing β, although convergence gets slightly slower. The right plot shows the corresponding MSE[2] evolution: adaptive diffusion is faster and the mismatch in steady-state is not noticeable. Figure 10 shows the adaptive weights λ_k for a few nodes. A weight close to one means that the corresponding node is performing better than its neighborhood's estimates, e.g., nodes 7 and 10. Likewise, nodes 2 and 5 were assigned small weights, meaning they are performing worse than the aggregated neighbors' estimates. Note how nodes with higher SNR, e.g., nodes 7 and 10, were assigned larger weights and nodes with lower SNR, e.g., nodes 2 and 5, were assigned smaller weights.

5. CONCLUDING REMARKS

We have proposed diffusion adaptive schemes to perform distributed estimation in a cooperative fashion. The schemes result in peer-to-peer algorithms suitable for general topologies and robust to link and node failures. Besides robustness and spatial diversity, diffusion protocols improve the network estimation performance. The diffusion techniques can also be extended to recursive least-squares formulations [3]. Analysis in the mean-square sense is available but will be approached in a complementary publication due to space constraints.

[2]Defined as $\xi_g(i) = \sum_{k=1}^{N} E|\boldsymbol{e}_k(i)|^2$.

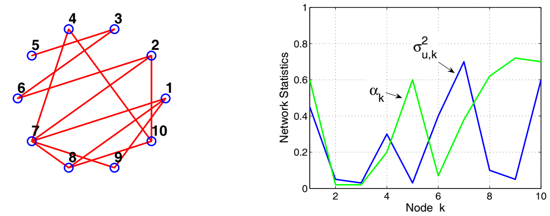

Fig. 8. Example 3: Network topology and statistical profile.

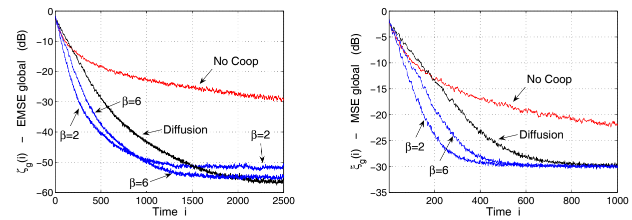

Fig. 9. Transient EMSE and MSE comparison.

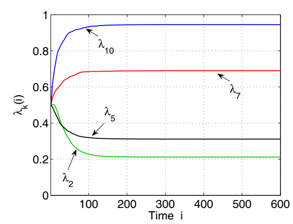

Fig. 10. The adaptive combiners.

6. REFERENCES

[1] C. Lopes and A. H. Sayed, "Distributed adaptive incremental strategies: formulation and performance analysis," *Proc. ICASSP*, vol.3, pp. 584–587, Toulouse, France, May 2006.

[2] C. Lopes and A. H. Sayed, "Distributed processing over adaptive networks," *Proc. Adaptive Sensor Array Processing Workshop*, MIT Lincoln Laboratory, MA, June 2006.

[3] A. H. Sayed and C. Lopes, "Distributed recursive least-squares strategies over adaptive networks," *Proc. 40th Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, October 2006.

[4] D. Estrin, G. Pottie and M. Srivastava. "Intrumenting the world with wireless sensor networks," *Proc. IEEE ICASSP*, pp. 2033-2036, Salt Lake City, UT, May 2001.

[5] A. H. Sayed. *Fundamentals of Adaptive Filtering*, Wiley, NJ, 2003.

[6] R. A. Horn and C. R. Johnson. *Matrix Analysis*, Cambridge, England, 1990.

[7] R. Bru, L. Elsner and M. Neumann "Convergence of infinite products of matrices and inner-outer iteration schemes" *Elec. Trans. on Numerical Analysis*, vol.2, pp.183-193, 1994.

[8] J. Arenas-Garcia, A. R. Figueiras-Vidal, and A. H. Sayed, "Mean-square performance of a convex combination of two adaptive filters," *IEEE Transactions on Signal Processing*, vol. 54, no. 3, pp. 1078–1090, March 2006.