Distributed Optimization via Diffusion Adaptation

Jianshu Chen, Sheng-Yuan Tu and Ali H. Sayed

Department of Electrical Engineering
University of California, Los Angeles, CA 90095

***Abstract*—We develop an iterative diffusion mechanism to optimize a global cost function in a distributed manner over a network of nodes. The cost function is assumed to consist of a collection of individual components, and diffusion strategy allows the nodes to cooperate and diffuse information in real-time. Compared to incremental methods, diffusion methods do not require the use of a cyclic path over the nodes and are more robust to node and link failure.**

***Index Terms*—Distributed optimization, diffusion adaptation, incremental strategy, learning, convex optimization.**

I. Introduction

We consider the problem of optimizing a global cost function in a distributed manner. The cost function is assumed to consist of the sum of individual components, and spatially distributed nodes are used to seek the common minimizer (or maximizer) through local interactions. There are already a couple of useful techniques for the solution of such optimization problems in a distributed manner [1]–[5]. Most notable among these methods is the incremental approach [2]–[5]. In this approach, a cyclic path is defined over the nodes and data are processed in a cyclic manner through the network until optimization is achieved. However, determining a cyclic path that covers all nodes is an NP-hard problem and, in addition, cyclic trajectories are prune to link and node failures. In earlier works [6]–[15], we introduced the concept of diffusion adaptation and showed how it can be used to solve global minimum mean-square-error estimation problems in real-time and in a distributed manner. In the diffusion approach, information is processed locally at the nodes and then diffused through a real-time sharing mechanism. This learning process can be used even in the presence of instantaneous approximations for the gradient vectors and helps reduce the effect of gradient noise on convergence. Besides, diffusion techniques are robust to node and link failure and do not require the use of a cyclic trajectory. In this article, we explain how the diffusion arguments of [13], [14] apply to the optimization of general cost functions in a distributed manner.

II. Problem Formulation

The objective is to determine an optimal $M \times 1$ vector w^o that minimizes a global cost function of the form:

$$J^{\text{glob}}(w) = \sum_{l=1}^{N} J_l(w) \tag{1}$$

where $J_l(w)$, $l = 1, 2, \ldots, N$, are the individual real-valued functions, assumed to be differentiable and convex. We assume $J^{\text{glob}}(w)$ in (1) is strictly convex so that the minimizer w^o is unique. In this article we consider the important case where each component function, $J_l(w)$, has a minimizer at the same w^o (see [16], [17] for examples in the context of biological networks). Our strategy to optimize $J^{\text{glob}}(w)$ in a distributed manner is based on three steps. First, using a second-order Taylor series expansion, we argue that the global cost function can be approximated by an alternative cost that is amenable to distributed optimization–see (9) further ahead. Second, each node optimizes the alternative cost via a steepest-descent procedure that relies on local data. Finally, the local estimates for w^o and gradient vectors are combined by each node and the procedure repeats itself in real-time.

To motivate the diffusion approach, we start by introducing a set of nonnegative coefficients $\{c_{l,k}\}$ that satisfy:

$$\sum_{k=1}^{N} c_{l,k} = 1, \quad c_{l,k} = 0 \text{ if } l \notin \mathcal{N}_k, \quad l = 1, 2, \ldots, N \tag{2}$$

where $\mathcal{N}_k$ denotes the neighborhood of node k. Using these coefficients, we can express $J^{\text{glob}}(w)$ from (1) as

$$J^{\text{glob}}(w) = J_k^{\text{loc}}(w) + \sum_{l \neq k}^{N} J_l^{\text{loc}}(w) \tag{3}$$

where

$$J_k^{\text{loc}}(w) = \sum_{l \in \mathcal{N}_k} c_{l,k} J_l(w) \tag{4}$$

In other words, for each node k, a new local cost function is introduced that corresponds to a weighted combination of the costs of its neighbors (including itself). Since the $\{c_{l,k}\}$ are all nonnegative and each $J_l(w)$ is convex, then $J_k^{\text{loc}}(w)$ is also a convex function.

Now assume there exists a w_l^{loc} that minimizes $J_l^{\text{loc}}(w)$; actually, the local cost function has a minimizer at the same w^o. Then, each $J_l^{\text{loc}}(w)$ can be approximated via a second-order Taylor series expansion as:

$$J_l^{\text{loc}}(w) \approx J_l^{\text{loc}}(w_l^{\text{loc}}) + \|w - w_l^{\text{loc}}\|_{\Gamma_l}^2 \tag{5}$$

where $\Gamma_l = \frac{1}{2}\nabla_w^2 J_l^{\text{loc}}(w_l^{\text{loc}})$ is the (scaled) Hessian matrix relative to w and evaluated at $w = w_l^{\text{loc}}$. Moreover, the notation $\|a\|_\Sigma^2$ denotes $a^T \Sigma a$ for any weighting matrix Σ. Substituting

Email: {jshchen, shinetu, sayed}@ee.ucla.edu. This work was supported in part by NSF grants CCF-1011918 and CCF-0942936.

(5) into the second term on the right-hand side of (3) gives:

$$J^{\text{glob}}(w) \approx J_k^{\text{loc}}(w) + \sum_{l \neq k} \|w - w_l^{\text{loc}}\|^2_{\Gamma_l} + \sum_{l \neq k} J_l^{\text{loc}}(w_l^{\text{loc}})$$

The last term in the above expression does not depend on w. Therefore, we can ignore this term so that optimizing $J^{\text{glob}}(w)$ is approximately equivalent to optimizing the following alternative general cost:

$$J^{\text{glob}'}(w) = J_k^{\text{loc}}(w) + \sum_{l \neq k} \|w - w_l^{\text{loc}}\|^2_{\Gamma_l} \tag{6}$$

III. Iterative Diffusion Solution

Expression (6) relates the desired global optimization problem to the newly-defined local cost function $J_k^{\text{loc}}(w)$. The relation is through the second term on the right-hand side of (6), which corresponds to a sum of quadratic terms involving the local estimates w_l^{loc}. Obviously, not all the local estimates w_l^{loc} are available at node k; only those estimates that originate from its neighbors can be assumed to be accessible by node k in a distributed solution. Likewise, not all the Hessian matrices Γ_l are available to node k. Nevertheless, expression (6) suggests a useful approximation that enables a powerful and elegant distributed solution.

Our first step is to replace the global cost $J^{\text{glob}'}(w)$ by reasonable localized approximations for it at every node k. Thus, initially we limit the summation on the right-hand side of (6) to the neighbors of node k and introduce

$$J_k^{\text{glob}'}(w) = J_k^{\text{loc}}(w) + \sum_{l \in \mathcal{N}_k \backslash \{k\}} \|w - w_l^{\text{loc}}\|^2_{\Gamma_l} \tag{7}$$

The cost (7) includes the quantities $\{w_l^{\text{loc}}, \Gamma_l\}$ that belong to the neighbors of node k. If desired, we can proceed with (7) and rely on the use of the Hessian matrices Γ_l (or approximations thereof) in the subsequent development. Nevertheless, in this paper, we simplify the argument in order to highlight the main ideas, and approximate each Γ_l by a multiple of the identity matrix, say, $\Gamma_l \approx b_{l,k} I_M$, for some nonnegative coefficients $\{b_{l,k}\}$. Such approximations are prevalent in stochastic approximation theory and they mark the difference between using a Newton's iterative method (which relies on the use of Hessian matrices and their inverses) or using a stochastic gradient method (where the Hessian matrix is approximated by a multiple of the identity matrix) (see [18, pp.142–147] and [19, pp.20–28]). Thus, we replace (7) by

$$J_k^{\text{glob}''}(w) = J_k^{\text{loc}}(w) \; + \sum_{l \in \mathcal{N}_k \backslash \{k\}} b_{l,k} \|w - w_l^{\text{loc}}\|^2 \tag{8}$$

As the derivation will show, we do not need to worry about how the scalars $\{b_{l,k}\}$ are selected. The argument so far suggests how to modify $J_k^{\text{loc}}(w)$ and replace it by the better approximation (8) for the global cost function (6). If we replace $J_k^{\text{loc}}(w)$ by its definition (5), we can rewrite (8) as

$$J_k^{\text{glob}''}(w) = \sum_{l \in \mathcal{N}_k} c_{l,k} J_l(w) + \sum_{l \in \mathcal{N}_k \backslash \{k\}} b_{l,k} \|w - w_l^{\text{loc}}\|^2 \tag{9}$$

Now, node k can apply a steepest-descent iteration to minimize $J_k^{\text{glob}''}(w)$ by using the (column) gradient vector:

$$\begin{aligned} w_{k,i} = w_{k,i-1} &- \mu_k \sum_{l \in \mathcal{N}_k} c_{l,k} \nabla_w J_l(w_{k,i-1}) \\ &- \mu_k \sum_{l \in \mathcal{N}_k \backslash \{k\}} 2 b_{l,k} (w_{k,i-1} - w_l^{\text{loc}}) \end{aligned} \tag{10}$$

The positive scalars $\{\mu_k\}$ denote step-size parameters. Among many other forms, we can implement (10) in two successive steps as follows:

$$\psi_{k,i} = w_{k,i-1} - \mu_k \sum_{l \in \mathcal{N}_k} c_{l,k} \nabla_w J_l(w_{k,i-1}) \tag{11}$$

$$w_{k,i} = \psi_{k,i} - \mu_k \sum_{l \in \mathcal{N}_k \backslash \{k\}} 2 b_{l,k} (w_{k,i-1} - w_l^{\text{loc}}) \tag{12}$$

Step (11) updates $w_{k,i-1}$ to an intermediate value $\psi_{k,i}$ by using local gradient vectors. Step (12) further updates $\psi_{k,i}$ to $w_{k,i}$. However, two issues arise while examining (12):

(a) First, iteration (12) requires knowledge of the local minimizers $\{w_l^{\text{loc}}\}$. The neighbors of node k do *not* know their local minimizers; each of these neighbors is actually performing steps similar to (11) and (12) to estimate their minimizers. This suggests that the readily available information about the $\{w_l^{\text{loc}}\}$ are the local estimates $\{\psi_{l,i}\}$. Therefore, we replace w_l^{loc} in (12) by $\psi_{l,i}$. This step helps diffuse information throughout the network.

(b) Second, the intermediate value $\psi_{k,i}$ is generally a better estimate for w^o than $w_{k,i-1}$ since it is obtained by incorporating information from the neighbors through (11). Therefore, we further replace $w_{k,i-1}$ in (12) by $\psi_{k,i}$. This step is reminiscent of incremental-type approaches to optimization, which have been widely studied in the literature [2]–[5].

With the substitutions described in (a) and (b) above, we arrive at the following Adapt-then-Combine (ATC) diffusion strategy (whose structure is the same as the ATC algorithm originally proposed in [8]–[14] for mean-square-error estimation):

$$\text{(ATC)} \quad \boxed{\begin{aligned} \psi_{k,i} &= w_{k,i-1} - \mu_k \sum_{l \in \mathcal{N}_k} c_{l,k} \nabla_w J_l(w_{k,i-1}) \\ w_{k,i} &= \sum_{l \in \mathcal{N}_k} a_{l,k} \psi_{l,i} \end{aligned}} \tag{13}$$

for some coefficients $\{a_{l,k}\}$ that satisfy the conditions:

$$\sum_{l=1}^{N} a_{l,k} = 1, \quad a_{l,k} = 0 \text{ if } l \notin \mathcal{N}_k \tag{14}$$

To run algorithm (13), we only need to select the coefficients $\{a_{l,k}, c_{l,k}\}$ satisfying (2) and (14); there is no need to worry about the intermediate coefficients $\{b_{l,k}\}$, which have been blended into the $\{a_{l,k}\}$. Similarly, if we reverse the order of steps (11) and (12), we can motivate the following alternative Combine-then-Adapt (CTA) diffusion strategy (whose structure is similar to the CTA algorithm originally proposed in

[6]–[14] for mean-square-error estimation; it was shown in [14] that ATC generally outperforms CTA):

$$\text{(CTA)} \quad \boxed{\begin{aligned} \psi_{k,i\text{-}1} &= \sum_{l\in\mathcal{N}_k} a_{l,k} w_{l,i\text{-}1} \\ w_{k,i} &= \psi_{k,i\text{-}1} - \mu_k \sum_{l\in\mathcal{N}_k} c_{l,k} \nabla_w J_l(\psi_{k,i\text{-}1}) \end{aligned}} \tag{15}$$

Adaptive diffusion strategies of the ATC and CTA types were first proposed in [6]–[14] and used to solve distributed minimum mean-square-error estimation problems over networks. A special case of the diffusion strategy (15) (corresponding to choosing $c_{l,k} = 0$ for $l \neq k$ and $c_{k,k} = 1$, i.e., without sharing gradient information) appeared later in the works [20], [21] and was used to solve distributed optimization problems that require all nodes to reach agreement. Diffusion recursions of the form (13) and (15) are more general in several respects. First, they do not only diffuse the local weight estimates, but they also diffuse the local gradient vectors. In other words, two complete sets of combination coefficients $\{a_{l,k}, c_{l,k}\}$ are used. Second, the combination weights $\{a_{l,k}\}$ are not required to be doubly stochastic (which means that the rows and columns of the corresponding weighting matrix $A = [a_{l,k}]$ should add up to one; as seen from (14), we only require the columns of A to add up to one). This condition will be shown further ahead to be sufficient to guarantee agreement when there is no noise in the data but, more importantly, the condition will *not* force nodes to seek agreement when data noise and gradient noise are present. The nodes will have the flexibility to tend to individual estimates that lie within a reasonable mean-square-error (MSE) performance bound from the optimal solution. Multi-agent systems in nature behave in this manner; they do not require exact agreement among their agents but allow for fluctuations due to individual levels of assessment and individual noise levels (see [14]–[17]). Finally, and importantly, the step-size parameters $\{\mu_k\}$ are not required to depend on the time index i and are not required to vanish as $i \to \infty$. Instead, they can assume constant values, which is critical to endow the network with continuous adaptation and learning abilities.

IV. Convergence Analysis

The arguments that led to (13) and (15) relied on some approximations that are typical of stochastic gradient and incremental approaches. The natural question now is to investigate the performance of the algorithms and to evaluate how well and how close they converge to w^o. In this paper, we study the convergence of (13) and (15) when noise is not present. In a related work [22], we derive expressions for the mean-square-deviation (MSD) of the algorithms in the presence of *noisy* gradient vectors; the MSD measures the mean of the squared error $\|w^o - w_{k,i}\|^2$ for every node k and in steady-state as $i \to \infty$.

A. Block Maximum Norm

For convergence analysis, we extend the argument of [15]. Let $x = \text{col}\{x_1, x_2, \ldots, x_N\} \in \mathbb{C}^{MN}$ denote a vector that is obtained by stacking N vectors of size $M \times 1$ each on top of each other. The block maximum norm of x is defined as

$$\|x\|_{b,\infty} \triangleq \max_{1\le k\le N} \|x_k\| \tag{16}$$

where $\|\cdot\|$ denotes the Euclidean norm of its vector argument. Furthermore, the induced block maximum norm of an $MN \times MN$ matrix B is defined as

$$\|B\|_{b,\infty} \triangleq \max_{x\in\mathbb{C}^{MN}, x\neq 0} \frac{\|Bx\|_{b,\infty}}{\|x\|_{b,\infty}} \tag{17}$$

We call upon the following lemma from [15].

Lemma 1: Let $Y = \text{diag}\{Y_1, \ldots, Y_N\}$ be an $MN \times MN$ block diagonal matrix with $M \times M$ unitary blocks $\{Y_k\}$, along its diagonal. Then, the following properties hold:

1) $\|Yx\|_{b,\infty} = \|x\|_{b,\infty}$, for all $x \in \mathbb{C}^{MN}$;
2) $\|YBY^*\|_{b,\infty} = \|B\|_{b,\infty}$, for all $B \in \mathbb{C}^{MN\times MN}$.

B. Convergence Analysis

We address the convergence behavior of both the ATC and CTA versions by viewing them as special cases of a more general diffusive algorithm of the following form:

$$\begin{cases} \phi_{k,i-1} = \sum_{l=1}^{N} p_{1,l,k} w_{l,i-1} \\ \psi_{k,i} = \phi_{k,i-1} - \mu_k \sum_{l=1}^{N} s_{l,k} \nabla_w J_l(\phi_{k,i-1}) \\ w_{k,i} = \sum_{l=1}^{N} p_{2,l,k} \psi_{l,i} \end{cases} \tag{18}$$

where the coefficients $\{p_{1,l,k}\}$, $\{s_{l,k}\}$, and $\{p_{2,l,k}\}$ are non-negative real coefficients corresponding to the $\{l,k\}$-th entries of matrices P_1, S, and P_2, respectively, which satisfy:

$$\boxed{\mathbb{1}^T P_1 = \mathbb{1}^T, \quad S\mathbb{1} = \mathbb{1}, \quad \mathbb{1}^T P_2 = \mathbb{1}^T} \tag{19}$$

Different choices for $\{P_1, P_2, S\}$ correspond to different cooperation modes. For example, the choice $P_1 = I$, $P_2 = A$, and $S = C$ corresponds to ATC, while the choice $P_1 = A$, $P_2 = I$, and $S = C$ corresponds to CTA. Introduce the error vectors:

$$\tilde{\phi}_{k,i} = w^o - \phi_{k,i}, \ \tilde{\psi}_{k,i} = w^o - \psi_{k,i}, \ \tilde{w}_{k,i} = w^o - w_{k,i}$$

Then, from (18), we have

$$\begin{cases} \tilde{\phi}_{k,i-1} = \sum_{l=1}^{N} p_{1,l,k} \tilde{w}_{l,i-1} \\ \tilde{\psi}_{k,i} = \tilde{\phi}_{k,i-1} + \mu_k \sum_{l=1}^{N} s_{l,k} \nabla_w J_l(\phi_{k,i-1}) \\ \tilde{w}_{k,i} = \sum_{l=1}^{N} p_{2,l,k} \tilde{\psi}_{l,i} \end{cases} \tag{20}$$

Using the fact that each component function $J_l(w)$ has a minimizer at the same w^o, and a result from [19, p.24], we can relate the gradient vectors in (20) to $\tilde{\phi}_{k,i-1}$ as follows:

$$\begin{aligned}\nabla_w J_l(\phi_{k,i-1}) &= -\left[\int_0^1 \nabla_w^2 J_l\big(w^o - t\tilde{\phi}_{k,i-1}\big)dt\right]\tilde{\phi}_{k,i-1}\\ &\triangleq -H_{l,k,i-1}\tilde{\phi}_{k,i-1} \qquad (21)\end{aligned}$$

The second equation in (20) can then be expressed as

$$\tilde{\psi}_{k,i} = \Big[I_M - \mu_k \sum_{l=1}^{N} s_{l,k} H_{l,k,i-1}\Big]\tilde{\phi}_{k,i-1} \qquad (22)$$

where $H_{l,k,i-1}$ depends on $\tilde{\phi}_{k,i-1}$. Introduce the global error vectors and matrices:

$$\tilde{\phi}_i = [\tilde{\phi}_{1,i} \cdots \tilde{\phi}_{N,i}]^T,\ \tilde{\psi}_i = [\tilde{\psi}_{1,i} \cdots \tilde{\psi}_{N,i}]^T,\ \tilde{w}_i = [\tilde{w}_{1,i} \cdots \tilde{w}_{N,i}]^T$$

$$\mathcal{P}_1 = P_1 \otimes I_M,\ \mathcal{P}_2 = P_2 \otimes I_M,\ \mathcal{S} = S \otimes I_M \qquad (23)$$

$$\mathcal{M} = \text{diag}\left\{\mu_1 I_M,\ \ldots,\ \mu_N I_M\right\} \qquad (24)$$

$$\mathcal{D}_{i-1} = \sum_{l=1}^{N} \text{diag}\Big\{s_{l,1}H_{l,1,i-1}, \cdots, s_{l,N}H_{l,N,i-1}\Big\} \qquad (25)$$

Then, recursions (20) and (22) lead to:

$$\boxed{\tilde{w}_i = \mathcal{P}_2^T[I_{MN} - \mathcal{M}\mathcal{D}_{i-1}]\mathcal{P}_1^T\tilde{w}_{i-1}} \qquad (26)$$

It follows that

$$\begin{aligned}\|\tilde{w}_i\|_{b,\infty} \le& \|\mathcal{P}_2^T\|_{b,\infty} \cdot \|I_{MN} - \mathcal{M}\mathcal{D}_{i-1}\|_{b,\infty}\\ &\cdot \|\mathcal{P}_1^T\|_{b,\infty} \cdot \|\tilde{w}_{i-1}\|_{b,\infty} \qquad (27)\end{aligned}$$

It was proved in [15] that $\|\mathcal{P}_1^T\|_{b,\infty}$ and $\|\mathcal{P}_2^T\|_{b,\infty}$ are bounded by one. Thus, it suffices to require

$$\sup_i \|I_{MN} - \mathcal{M}\mathcal{D}_{i-1}\|_{b,\infty} < 1 \qquad (28)$$

in order to ensure that the error $\tilde{w}_i$ converges to zero. The following theorem gives a condition for (28) to hold.

Theorem 1: Suppose the Hessian matrices satisfy

$$\lambda_{l,\min} I_M \le \nabla_w^2 J_l(w) \le \lambda_{l,\max} I_M, \quad l = 1, 2, \ldots, N \quad (29)$$

with $\sum_{l=1}^{N} s_{l,k}\lambda_{l,\min} > 0$. Then, condition (28) holds for step-sizes that satisfy:

$$\boxed{0 \le \mu_k \le 2\Big(\sum_{l=1}^{N} s_{l,k}\lambda_{l,\max}\Big)^{-1}, \quad k = 1, \ldots, N} \qquad (30)$$

Proof: The idea is to diagonalize the block diagonal matrix $\mathcal{D}_{i-1}$ by using a block unitary matrix transform. Then, the result would follow from Lemma 1. Details are omitted. ■

To illustrate the performance of the algorithms through a numerical example, Table I shows the values obtained for the network mean-square-deviation (MSD) defined as

$$\text{MSD}^{\text{network}} = \frac{1}{N}\sum_{k=1}^{N}\|w^o - w_{k,\infty}\|^2 \qquad (31)$$

TABLE I
MEAN-SQUARE-DEVIATION (MSD) OF THE DIFFUSION STRATEGIES

Algorithm	ATC	CTA
MSD	1.44×10^{-31}	2.63×10^{-31}

The MSD values were computed by applying 200 iterations of the ATC and CTA algorithms to a 10-node random network to minimize the following (localization) cost function over w:

$$J^{\text{glob}}(w) = \sum_{k=1}^{N}\left|d_k - \|w - x_k\|^2\right|^2 \qquad (32)$$

where each node was assumed to know its location x_k and its distance d_k to the target at w^o. The step-size was set to $\mu_k = 0.02$.

REFERENCES

[1] J. N. Tsitsiklis and M. Athans, "Convergence and asymptotic agreement in distributed decision problems," *IEEE Trans. Autom. Control*, vol. 29, no. 1, pp. 42–50, 1984.

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

[3] A. Nedic and D. P. Bertsekas, "Incremental subgradient methods for nondifferentiable optimization," *SIAM J. Optim.*, vol. 12, no. 1, pp. 109–138, 2001.

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

[5] 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.

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

[7] A. H. Sayed and C. G. Lopes, "Adaptive processing over distributed networks," *IEICE Trans. Fund. of Electron., Commun. and Comput. Sci.*, vol. E90-A, no. 8, pp. 1504–1510, 2007.

[8] F. S. Cattivelli, C. G. Lopes, and A. H. Sayed, "A diffusion RLS scheme for distributed estimation over adaptive networks," in *Proc. IEEE Workshop on Signal Process. Advances Wireless Comm. (SPAWC)*, Helsinki, Finland, June 2007, pp. 1–5.

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

[10] F. S. Cattivelli, C. G. Lopes, and A. H. Sayed, "Diffusion strategies for distributed Kalman filtering: Formulation and performance analysis," in *Proc. IAPR Workshop on Cognitive Inf. Process. (CIP)*, Santorini, Greece, June 2008, pp. 36–41.

[11] 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, July 2008.

[12] F. S. Cattivelli and A. H. Sayed, "Diffusion mechanisms for fixed-point distributed Kalman smoothing," in *Proc. EUSIPCO*, Lausanne, Switzerland, Aug. 2008, pp. 1–4.

[13] F. S. Cattivelli and A. H. Sayed, "Diffusion LMS algorithms with information exchange," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, Pacific Grove, CA, Nov. 2008, pp. 251–255.

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

[15] N. Takahashi, I. Yamada, and A. H. Sayed, "Diffusion least-mean squares with adaptive combiners: formulation and performance analysis," *IEEE Trans. Signal Process.*, vol. 58, no. 9, pp. 4795–4810, Sep. 2010.

[16] F. S. Cattivelli and A. H. Sayed, "Modeling bird flight formations using diffusion adaptation," *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 2038–2051, May 2011.

[17] S.-Y. Tu and A. H. Sayed, "Mobile adaptive networks," *IEEE J. Sel. Topics. Signal Process.*, vol. 5, no. 4, pp. 649–664, Aug. 2011.

[18] A. H. Sayed, *Adaptive Filters*, Wiley, NJ, 2008.

[19] B. Polyak, *Introduction to Optimization*, Optimization Software, NY, 1987.

[20] S. S. Ram, A. Nedic, and V. V. Veeravalli, "Distributed stochastic subgradient projection algorithms for convex optimization," *J. Optim. Theory Appl.*, vol. 147, no. 3, pp. 516–545, 2010.

[21] P. Bianchi, G. Fort, W. Hachem, and J. Jakubowicz, "Convergence of a distributed parameter estimator for sensor networks with local averaging of the estimates," in *Proc. IEEE ICASSP*, May 2011, pp. 3764–3767.

[22] J. Chen and A. H. Sayed, "Performance of diffusion adaptation for collaborative optimization," *submitted for publication*.