EUSIPCO 2013 1569744697

ON THE BENEFITS OF DIFFUSION COOPERATION FOR DISTRIBUTED OPTIMIZATION AND LEARNING

Jianshu Chen and Ali H. Sayed

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

ABSTRACT

This work characterizes the nature of the limit point of distributed strategies for adaptation and learning over networks in the general case when the combination policy is not necessarily doubly stochastic and when the individual risks do not necessarily share a common minimizer. It is shown that, for sufficiently small step-sizes, the limiting behavior of the network is mainly influenced by the right-eigenvector of the combination policy corresponding to the single eigenvalue at one. It is also shown that the limit point of the network is the unique solution to a certain fixed-point equation determined by the entries of this eigenvector. The arguments show further that even when only partial information is available to the agents, cooperation over a connected network enables the agents to attain the same level of performance as a centralized solution.

Index Terms— Distributed optimization, diffusion strategy, consensus strategy, Pareto optimality.

1. INTRODUCTION

We examine two important classes of strategies for the optimization of aggregate cost functions by a connected network of N distributed agents. One class of strategies relies on the consensus implementation [1–8] and the second class of strategies relies on the diffusion implementation [9–14]. In both implementations, the agents use stochastic gradient recursions and collaborate locally to estimate some parameter of interest. The step-sizes used by the agents can be identical or different. The algorithms can be motivated as follows. Let $J^{\rm glob}(w)$ denote the aggregate cost function

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

which is defined in terms of individual cost functions $J_k(w)$ assigned to each agent, for $k = 1, \ldots, N$. The objective is to minimize $J^{\rm glob}(w)$, which is assumed to be strongly convex.

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

There are several variants of distributed strategies. Consensus and the adapt-then-combine (ATC) diffusion strategies pursue the above objective by employing recursions of the following form:

$$\begin{cases} \boldsymbol{\psi}_{k,i-1} = \displaystyle\sum_{\ell\in\mathcal{N}_k} a_{\ell k}\boldsymbol{w}_{\ell,i-1} \\ \boldsymbol{w}_{k,i} = \boldsymbol{\psi}_{k,i-1} - \mu_k\widehat{\nabla_w J_k}(\boldsymbol{w}_{k,i-1}) \end{cases} \quad \text{[consensus]} \tag{2}$$

and

$$\begin{cases} \boldsymbol{\psi}_{k,i} = \boldsymbol{w}_{k,i-1} - \mu_k\widehat{\nabla_w J_k}(\boldsymbol{w}_{k,i-1}) \\ \boldsymbol{w}_{k,i} = \displaystyle\sum_{\ell\in\mathcal{N}_k} a_{\ell k}\boldsymbol{\psi}_{\ell,i} \end{cases} \quad \text{[diffusion]} \tag{3}$$

where the μ_k are non-negative step-size parameters, $\boldsymbol{w}_{k,i}$ denotes the estimate computed by agent k at time i, and $\widehat{\nabla_w J_k}(\cdot)$ represents a stochastic approximation for the true gradient vector of $J_k(w)$ with respect to w. We require at least one $\mu_k > 0$ so that at least one agent in the network is performing adaptation. Some of the other step-sizes can be zero, in which case the corresponding agents would only be participating in the aggregation of information. Moreover, the symbol $\mathcal{N}_k$ in (2)–(3) denotes the set of neighbors of agent k. The $\{a_{\ell k}\}$ are non-negative combination coefficients that satisfy

$$\sum_{\ell\in\mathcal{N}_k} a_{\ell k} = 1, \quad a_{\ell k} = 0 \text{ if } \ell \notin \mathcal{N}_k \tag{4}$$

If we introduce the $N \times N$ matrix $A = [a_{\ell k}]$, then condition (4) implies that A is a left-stochastic matrix, namely, it satisfies $A^T\mathbb{1} = \mathbb{1}$, where $\mathbb{1}$ denotes the vector with all entries equal to one.

In the diffusion implementation (3), the variable $\boldsymbol{\psi}_{\ell,i}$ is an intermediate estimator that is shared among the neighbors in lieu of the estimators $\boldsymbol{w}_{\ell,i-1}$ used by the consensus implementation (2). In doing so, diffusion allows information to spread more thoroughly through the network. It is clear from (2) and (3) that the consensus and diffusion implementations have exactly the same computational complexity, and yet it was shown in [15] that the diffusion dynamics leads to improved mean-square-error performance. Diminishing step-sizes of the form $\mu_k(i) = c_k/(i+1)$ for some constants

$c_k > 0$, can also be used in (2)–(3). Nevertheless, we focus in this work on the case of constant step-sizes in order to endow the agents with continuous learning and adaptation abilities. For studies involving diminishing step-sizes, readers are referred to [2–5, 16].

One important question is the following. The above recursive schemes are motivated as distributed solutions for the minimization of the aggregate cost function (1). When the individual costs are all minimized at the *same* location, w^o, or when the combination matrix A happens to be doubly stochastic (i.e., each of its columns and each of its rows add up to one), then results in the literature [5, 6, 11, 12] already establish that the estimates $\boldsymbol{w}_{k,i}$ generated by both strategies (2)–(3) converge within small mean-square-error to the unique minimizer w^o of (1). But what if A is not doubly stochastic and what if the individual costs $J_k(w)$ do not share a common minimizer? Where do these strategies converge to? These questions are relevant because it is known that the optimal combination policy that minimizes the steady-state mean-square-error is not necessarily doubly stochastic [17] — see also expression (36) further ahead.

Building on recent results from [13], this article answers the above question in an interesting way. The conclusion will establish that, for sufficiently small step-sizes, the distributed schemes (2)–(3) do not converge to the minimizer of (1) but rather to the unique solution of the following equation:

$$\sum_{k=1}^{N} p_k \nabla_w J_k(w) \;=\; 0 \tag{5}$$

where the $\{p_k\}$ denote positive scalars that will be constructed later from the step-sizes and from the entries of the right-eigenvector of A corresponding to the eigenvalue at one — see (13). Moreover, the analysis will show that the solution of (5) can be interpreted as being a Pareto optimal solution to a multi-objective optimization problem. The analysis will further show that when A is doubly stochastic or when the $\{J_k(w)\}$ have a common minimizer, then the solution of (5) agrees with the minimizer of (1). In this way, equation (5) can be viewed as the desired relation that characterizes the limit point of the network for all scenarios. These conclusions are significant for various reasons:

(a) They show how the choice of the combination policy A influences the limit point of the learning process (through the scalars $\{p_k\}$).

(b) They provide a way to select A in order to drive the network towards a desired limit point, including the limit point of the original optimization problem (1). For this latter purpose, we would simply need to ensure that the resulting $\{p_k\}$ are all identical.

(c) They allow us to quantify the benefit of cooperation, as the results in the following sections reveal.

2. PROBLEM FORMULATION

Thus, consider a network of N connected agents, where each agent k receives a stream of data $\{\boldsymbol{x}_{k,i}\}$ arising from some underlying distribution. In this paper, we consider the case where the topology is static over time and the communication links are noise free. The networked multi-agent system would then like to extract some useful information about the underlying process from the distributed data. To measure the quality of the inference task, an individual convex cost function, $J_k(w)$, is associated with each agent k, where w denotes an $M \times 1$ parameter vector. The agents are interested in minimizing an aggregate cost function of the form (1). Based on whether the individual costs $\{J_k(w)\}$ share a common minimizer or not, we can classify problems of the form (1) into two broad categories.

2.1. Category I: Distributed Learning

In this case, the data streams $\{\boldsymbol{x}_{k,i}\}$ are assumed to be generated by (possibly different) distributions that depend on the same parameter vector $w^o \in \mathbb{R}^M$. The objective is then to estimate this common parameter w^o in a distributed manner. To do so, we first need to associate with each agent k a cost function $J_k(w)$ that measures how well some arbitrary parameter w approximates w^o. The cost $J_k(w)$ should be such that w^o is one of its minimizers. More formally, let $\mathcal{W}_k^o$ denote the set of vectors that minimize the selected $J_k(w)$, then

$$w^o \in \mathcal{W}_k^o \triangleq \Big\{ w : \arg\min_w J_k(w) \Big\} \tag{6}$$

for $k = 1, \ldots, N$. Since $J^{\rm glob}(w)$ is assumed to be strongly convex, then the intersection of the sets $\mathcal{W}_k^o$ should contain the single element w^o:

$$w^o \in \mathcal{W}^o \triangleq \bigcap_{k=1}^{N} \mathcal{W}_k^o \tag{7}$$

The main motivation for cooperation in this case is that the data collected at each agent k may not be sufficient to uniquely identify w^o since w^o is not the unique element in $\mathcal{W}_k^o$; this happens, for example, when the individual costs $J_k(w)$ are not necessarily *strictly* convex. However, once the individual costs are aggregated into (1) and the aggregate function is strongly convex, then w^o is the unique element in $\mathcal{W}^o$. In this way, the cooperative minimization of $J^{\rm glob}(w)$ allows the agents to estimate w^o. In addition, it is generally beneficial to have more agents involved in the learning process to reduce the effect of statistical perturbations in the data [17].

2.2. Category II: Distributed Optimization

In this case, we include situations where the individual costs $J_k(w)$ do not have a common minimizer, i.e., $\mathcal{W}^o = \emptyset$. The

optimization problem should then be viewed as one of solving a multi-objective minimization problem of the form:

$$\min_{w} \{J_1(w), \ldots, J_N(w)\} \tag{8}$$

A vector w^o is said to be a Pareto optimal solution to (8) if there does not exist another vector w that is able to improve (i.e., reduce) any individual cost without degrading (increasing) some of the other costs. Pareto optimal solutions are not necessarily unique. It is known that the problem of determining Pareto optimal solutions can be transformed into the optimization of cost functions of a form similar to (1) by means of a scalarization technique [12, 18]. Specifically, we replace (1) by

$$J^{\text{glob}}(w) = \sum_{k=1}^{N} \pi_k J_k(w) \tag{9}$$

where the $\{\pi_k\}$ are positive weighting coefficients that we are free to choose. Then, each set of coefficients $\{\pi_k\}$ leads to a Pareto optimal solution w^o to problem (8). Since we can rescale the individual costs in (9) above as

$$J_k(w) \;\longleftarrow\; \pi_k J_k(w) \tag{10}$$

then we are reduced again to the scenario described by (1) and the same recursions (2)–(3) continue to be applicable.

The question we would like to address now is the following: given individual costs $\{J_k(w)\}$ and a combination policy A, what is the limit point of the distributed strategies (2) or (3)?

3. LIMIT POINT OF LEARNING PROCESS

3.1. Diffusion and Consensus Strategies

We observe from (2)–(3) that there are two types of learning processes involved in the dynamics of each agent k: (i) self-learning with stochastic gradients $\{\widehat{\nabla_w J_k}(\cdot)\}$ from locally sensed data and (ii) social learning using combination steps from neighbors. All nodes implement the same self- and social learning structure. As a result, the learning dynamics of all nodes in the network are coupled; knowledge exploited from local data at node k will be propagated to its neighbors and from there to their neighbors in a diffusive learning process.

We introduce a couple of assumptions that are sufficient to guarantee the convergence of the learning process.

Assumption 1 (Strongly connected network)**.** *We require A to be a primitive left-stochastic matrix, i.e., $A^T\mathbb{1} = \mathbb{1}$ and there exists a finite positive integer j_0 such that all entries of A^{j_0} are strictly positive.* ■

Assumption 1 is automatically satisfied if the network is connected and there is at least one $a_{kk} > 0$ for some node k. It then follows from the Perron-Frobenius Theorem [19] that the matrix A has an eigenvalue at one with multiplicity one and that all other eigenvalues of A are strictly less than one in magnitude. Obviously, $\mathbb{1}^T$ is a left eigenvector for A corresponding to the eigenvalue at one. Let θ denote the right eigenvector corresponding to the eigenvalue at one and whose entries are normalized to add up to one, i.e.,

$$A\theta = \theta \quad \text{and} \quad \mathbb{1}^T\theta = 1 \tag{11}$$

Then, the Peron-Frobenius Theorem further ensures that all entries of θ are positive.

Definition 1 (p-vector)**.** *Let*

$$\mu_{\max} \triangleq \max_k \mu_k \tag{12}$$

so that $\mu_k = \mu_{\max}\beta_k$ for some nonnegative scalars $0 \le \beta_k \le 1$. We define

$$p \triangleq \text{col}\{\theta_1\beta_1, \ldots, \theta_N\beta_N\} \tag{13}$$

where θ_k is the kth entry of the vector θ.

The vector p plays an important role in characterizing the limit point and the steady-state mean-square-error performance of the distributed strategies (2) and (3).

Furthermore, we denote the difference between the true and approximate gradient vectors as the gradient noise $\boldsymbol{v}_{k,i}(\cdot)$:

$$\widehat{\nabla_w J_k}(\boldsymbol{w}) = \nabla_w J_k(\boldsymbol{w}) + \boldsymbol{v}_{k,i}(\boldsymbol{w}) \tag{14}$$

Assumption 2 (Gradient noise)**.** *There exist $\alpha_k \ge 0$ and $\sigma_{v,k}^2 \ge 0$ such that, for all $\boldsymbol{w} \in \mathcal{F}_{i-1}$:*

$$\mathbb{E}\{\boldsymbol{v}_{k,i}(\boldsymbol{w}) \mid \mathcal{F}_{i-1}\} = 0 \tag{15}$$

$$\mathbb{E}\left\{\|\boldsymbol{v}_{k,i}(\boldsymbol{w})\|^2\right\} \le \alpha_k \cdot \mathbb{E}\|\nabla_w J_k(\boldsymbol{w})\|^2 + \sigma_{v,k}^2 \tag{16}$$

for all i, k, where $\mathcal{F}_{i-1}$ denotes the past history of all iterates $\{\boldsymbol{w}_{k,j}\}$ up to time $i-1$ for all k. ■

The above assumption is standard in stochastic approximation theory [20] and we explained why it is necessary in the context of adaptive solutions in [11].

Finally, we require the true gradient vectors $\{\nabla_w J_k(w)\}$ to satisfy the following conditions.

Assumption 3 (Lipschitz gradients)**.** *There exist $\lambda_U \ge 0$ such that for all $x, y \in \mathbb{R}^M$ and all k:*

$$\|\nabla_w J_k(x) - \nabla_w J_k(y)\| \le \lambda_U \cdot \|x - y\| \tag{17}$$

where the subscript "U" in λ_U refers to an upper bound. ■

Assumption 4 (Global observability)**.** *There exists $\lambda_L > 0$ such that for all $w \in \mathbb{R}^M$:*

$$\sum_{k=1}^{N} p_k \nabla_w^2 J_k(w) \ge \lambda_L I_M \tag{18}$$

where the subscript "L" in λ_L refers to a lower bound, and the $\{p_k\}$ denote entries of the vector p. ■

3.2. Main Result

We summarize the main result in the following theorem; its proof follows from the arguments used in [13].

Theorem 1 (Limit point and performance)**.** *Suppose Assumptions 1–4 hold. Then, there exists a unique solution $w^o \in \mathbb{R}^M$ to the following equation:*

$$\sum_{k=1}^{N} p_k \nabla_w J_k(w) = 0 \tag{19}$$

Moreover, for sufficiently small step-sizes, both the consensus strategy (2) *and the diffusion strategy* (3) *will converge to this w^o in the mean-square sense at the following rate:*

$$r \approx 1 - 2\mu_{\max}\lambda_{\min}(R_c) \tag{20}$$

with the steady-state mean-square-error (MSE) at each agent approximated by:

$$\lim_{i\to\infty} \mathbb{E}\|\tilde{\boldsymbol{w}}_{k,i}\|^2 \approx \frac{\mu_{\max}}{2} \cdot \mathrm{Tr}\left\{(p^T \otimes I_M)\mathcal{R}_v(p \otimes I_M)R_c^{-1}\right\} \tag{21}$$

where $\mu_{\max}$ is defined in (12) *and*

$$R_c \triangleq \sum_{k=1}^{N} p_k H_k \tag{22}$$

$$H_k \triangleq \nabla_w^2 J_k(w^o) \tag{23}$$

$$\mathcal{R}_v \triangleq \lim_{i\to\infty} \mathbb{E}\left\{\boldsymbol{v}_i(w^o)\boldsymbol{v}_i(w^o)^T\right\} \tag{24}$$

$$\boldsymbol{v}_i(w^o) \triangleq \mathrm{col}\left\{\boldsymbol{v}_{1,i}(w^o), \ldots, \boldsymbol{v}_{N,i}(w^o)\right\} \tag{25}$$

■

One important observation from (21) is that the performance is independent of the agent index; different agents will have almost the same performance to first order in the step-sizes. In fact, the convergence rate (20) and the MSE performance level (21) have the same values that would result from the following centralized solution:

$$\boldsymbol{w}_{\mathrm{cent},i} = \boldsymbol{w}_{\mathrm{cent},i-1} - \mu_{\max} \sum_{k=1}^{N} p_k \widehat{\nabla_w J_k}(\boldsymbol{w}_{\mathrm{cent},i-1}) \tag{26}$$

where $\boldsymbol{w}_{\mathrm{cent},i}$ denotes the estimate generated by the centralized recursion. In [17], results similar to (20)–(21) were derived for the special case of diffusion-LMS adaptive networks, where the Hessian matrices and step-sizes were assumed to be the same across all agents.

4. BENEFITS OF COOPERATION

4.1. Distributed Learning

4.1.1. Working under Partial Observation

Under the scenario described by (7), the solution of (19) agrees with the unique minimizer w^o for $J^{\mathrm{glob}}(w)$ given by (1) regardless of the $\{p_k\}$ and, therefore, regardless of the combination policy A. Therefore, Theorem 1 ensures that the estimator $\boldsymbol{w}_{k,i}$ at each agent k converges to this unique w^o at a centralized rate and MSE performance. Note that Assumption 4 can be satisfied without requiring each $J_k(w)$ to be strongly convex. Instead, we only require $J^{\mathrm{glob}}(w)$ to be strongly convex. In other words, we do not need each agent to have complete information about w^o; we only need the network to have enough information to determine w^o uniquely. Although the individual agents in this case have partial information about w^o, the distributed strategies (2) and (3) enable them to attain the same performance level as a centralized solution. The following example illustrates the idea in the context of distributed LMS estimation.

Example. *Suppose each agent k collects data $\{\boldsymbol{u}_{k,i}, \boldsymbol{d}_k(i)\}$ that are related via the linear model:*

$$\boldsymbol{d}_k(i) = \boldsymbol{u}_{k,i}w^o + \boldsymbol{z}_{k,i} \tag{27}$$

where $\boldsymbol{z}_{k,i}$ is a zero mean noise process, and the regressor $\boldsymbol{u}_{k,i}$ is $1 \times M$. The objective is to estimate w^o in a distributed manner by minimizing (1) *where*

$$J_k(w) = \mathbb{E}|\boldsymbol{d}_k(i) - \boldsymbol{u}_{k,i}w|^2 \tag{28}$$

When the covariance matrix $R_{u,k} \triangleq \mathbb{E}[\boldsymbol{u}_{k,i}^T\boldsymbol{u}_{k,i}]$ is rank deficient, then $J_k(w)$ would not be strongly convex and there would be infinitely many minimizers to $J_k(w)$. In this case, the information provided to agent k via (27) *is not sufficient to determine w^o uniquely. However, if the global cost function is strongly convex, which can be shown to be equivalent to requiring:*

$$\sum_{k=1}^{N} p_k R_{u,k} > \lambda_L I_M \tag{29}$$

then the information collected over the entire network is rich enough to learn the unique w^o. As long as (29) *holds for one set of positive $\{p_k\}$, it will hold for all other positive $\{p_k\}$. A "network observability" condition similar to* (29) *was used in [3] to characterize the sufficiency of information over the network in a similar context of distributed estimation over linear models albeit with diminishing step-sizes.* ■

4.1.2. Optimizing the MSE Performance

Since the distributed strategies (2) and (3) converge to the minimizer w^o of (1) for any set of $\{p_k\}$, we can then consider selecting the $\{p_k\}$ to optimize the MSE performance. Consider the case where $H_k \equiv H$ and $\mu_k \equiv \mu$ and assume the gradient noises are asymptotically uncorrelated across the agents so that $\mathcal{R}_v$ becomes block diagonal with entries denoted by:

$$\mathcal{R}_v = \mathrm{diag}\{R_{v,1}, \ldots, R_{v,N}\} \tag{30}$$

Then, we have $\beta_k = 1$ and $p_k = \theta_k$ in which case expressions (20)–(21) become

$$r \approx 1 - 2\mu\lambda_{\min}(H) \tag{31}$$

$$\lim_{i\to\infty} \mathbb{E}\|\tilde{\boldsymbol{w}}_{k,i}\|^2 \approx \frac{\mu}{2} \cdot \sum_{k=1}^{N} \theta_k^2 \mathrm{Tr}\left(R_{v,k} H^{-1}\right) \tag{32}$$

The optimal positive coefficients $\{\theta_k\}$ that minimize (32) subject to $\sum_{k=1}^{N} \theta_k = 1$ is given by

$$\theta_k^o = \frac{[\mathrm{Tr}(R_{v,k} H^{-1})]^{-1}}{\sum_{\ell=1}^{N} [\mathrm{Tr}(R_{v,\ell} H^{-1})]^{-1}} \tag{33}$$

and the optimal MSE is

$$\mathrm{MSE}^{\mathrm{opt}} \approx \frac{\mu}{2} \cdot \left[\sum_{\ell=1}^{N} \frac{1}{\mathrm{Tr}(R_{v,\ell} H^{-1})}\right]^{-1} \tag{34}$$

The optimal right-eigenvector $\theta^o = \mathrm{col}\{\theta_1^o, \ldots, \theta_N^o\}$ can be implemented by selecting the combination policy A as the following Hasting's rule [17, 21, 22]:

$$a_{\ell k}^o = \begin{cases} \dfrac{(\theta_k^o)^{-1}}{\max\{|\mathcal{N}_k| \cdot (\theta_k^o)^{-1}, |\mathcal{N}_\ell| \cdot (\theta_\ell^o)^{-1}\}}, & \ell \in \mathcal{N}_k \backslash \{k\} \\ 1 - \displaystyle\sum_{m \in \mathcal{N}_k \backslash \{k\}} a_{mk}^o, & \ell = k \end{cases} \tag{35}$$

4.1.3. Left-Stochastic Combination Policies

If we choose a doubly-stochastic combination policy for A, then $\theta_k = 1/N$. Note from (31) that the convergence rate would be the same for both the left-stochastic policy (35) and the doubly-stochastic policy. However, the steady-state MSE for the doubly stochastic policy is worse than the performance by the optimal left-stochastic policy:

$$\mathrm{MSE}^{\mathrm{ds}} = \frac{\mu}{2} \cdot \frac{1}{N^2} \sum_{k=1}^{N} \mathrm{Tr}\left(R_{v,k} H^{-1}\right) \geq \mathrm{MSE}^{\mathrm{opt}} \tag{36}$$

where equality holds only when all agents have the same noise covariance matrices, i.e., $R_{v,\ell} \equiv R_v$.

4.2. Distributed Pareto Optimization

When the individual costs $\{J_k(w)\}$ do not necessarily share a common minimizer, then each set of coefficients $\{p_k\}$ will lead to a different minimizer of the fixed point equation (19). This minimizer will correspond to a Pareto optimal solution. Again, if we assume that $H_k \equiv H$, $\mu_k \equiv \mu$, and that the gradient noises are asymptoically uncorrelated as in (30), then we will obtain the same results along the lines of (31)–(32). Therefore, if we want to reach a Pareto-optimal solution w^o that has the smallest steady-state MSE, then we should choose the same Hasting's rule as (35). Readers may consult [12] for examples.

5. REFERENCES

[1] D.P. Bertsekas and J.N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Athena Scientific, Belmont, 1997.

[2] J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *IEEE Trans. Autom. Control*, vol. 31, no. 9, pp. 803–812, 1986.

[3] S. Kar and J. M. F. Moura, "Convergence rate analysis of distributed gossip (linear parameter) estimation: Fundamental limits and tradeoffs," *IEEE J. Sel. Topics. Signal Process.*, vol. 5, no. 4, pp. 674–690, Aug. 2011.

[4] 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.

[5] 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.

[6] A. Nedic and A. Ozdaglar, "Cooperative distributed multi-agent optimization," *Convex Optimization in Signal Processing and Communications,* Y. Eldar and D. Palomar (Eds.), Cambridge University Press, pp. 340–386, 2010.

[7] A. G. Dimakis, S. Kar, J. M. F. Moura, M. G. Rabbat, and A. Scaglione, "Gossip algorithms for distributed signal processing," *Proc. IEEE*, vol. 98, no. 11, pp. 1847–1864, Nov. 2010.

[8] S. Barbarossa and G. Scutari, "Bio-inspired sensor network design," *IEEE Signal Process. Mag.*, vol. 24, no. 3, pp. 26–35, May 2007.

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

[10] 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.

[11] 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.

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

[13] J. Chen and A. H. Sayed, "On the limiting behavior of distributed optimization strategies," in *Proc. 50th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, Oct. 2012, pp. 1535–1542.

[14] S. Chouvardas, K. Slavakis, and S. Theodoridis, "Adaptive robust distributed learning in diffusion sensor networks," *IEEE Trans. Signal Process.*, vol. 59, no. 10, pp. 4692–4707, Oct. 2011.

[15] 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.

[16] 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*, Prague, Czech, May 2011, pp. 3764–3767.

[17] X. Zhao and A. H. Sayed, "Performance limits for distributed estimation over LMS adaptive networks," *IEEE Trans. Signal Process.*, vol. 60, no. 10, pp. 5107–5124, Oct. 2012.

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

[19] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, 1990.

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

[21] S. Boyd, P. Diaconis, and L. Xiao, "Fastest mixing Markov chain on a graph," *SIAM Rev.*, vol. 46, no. 4, pp. 667–689, Dec. 2004.

[22] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, vol. 57, no. 1, pp. 97–109, Apr. 1970.