# On the Probability Distribution of Distributed Optimization Strategies

Jianshu Chen and Ali H. Sayed

Department of Electrical Engineering
University of California, Los Angeles, CA 90034

***Abstract*—We study the steady-state probability distribution of diffusion and consensus strategies that employ constant step-sizes to enable continuous adaptation and learning. We show that, in the small step-size regime, the estimation error at each agent approaches a Gaussian distribution. More importantly, the covariance matrix of this distribution is shown to coincide with the error covariance matrix that would result from a centralized stochastic-gradient strategy. The results hold regardless of the connected topology and help clarify the convergence and learning behavior of distributed strategies in an interesting way.**

***Index Terms*—Diffusion strategy, consensus strategy, distributed stochastic optimization, central limit theorem, steady-state performance**

## I. Introduction

In multi-agent systems, agents interact with each other to solve a problem of common interest, such as an optimization problem in a distributed manner. Two useful strategies that can be used to guide the interactions of agents over a network are consensus strategies [1]–[6] and diffusion strategies [7]–[11].

We assume the strategies employ constant step-sizes, as opposed to decaying step-sizes, in order to enable continuous learning and adaptation. Under this condition, earlier results [10]–[12] derived closed-form expressions to characterize in some detail the rate of convergence and the steady-state mean-square-error (MSE) performance of the distributed strategies over connected networks. The MSE expressions are useful because they reveal the *expected* behavior of the learning process over repeated experiments. In this work, we focus instead on examining the behavior of a *single* realization of the learning curve. This is an important objective because, in practice, one usually runs a distributed strategy once and would like to know what performance guarantees can be expected with high probability under this scenario.

It was observed earlier in [13] that there is an interesting interplay between the mean-square behavior of a stand-alone adaptive agent and its almost-sure behavior. Investigating a similar issue over networks is substantially more demanding due to the coupling among the agents. We shall examine single-realization behavior for networks by studying the probability that the steady-state solutions at the agents, $\{\boldsymbol{w}_{k,i}\}$ as $i \to \infty$, deviate away from the desired solution $w^o$. One way to bound this probability is to use the available MSE results and to invoke Chebyshev's inequality to write for $i \gg 1$ and for any $\epsilon > 0$:

$$\Pr\left\{\|\boldsymbol{w}_{k,i} - w^o\| > \epsilon\right\} \leq \frac{\mathbb{E}\|\boldsymbol{w}_{k,i} - w^o\|^2}{\epsilon^2} \qquad (1)$$

However, it is well-known that the Chebyshev inequality generally leads to a loose bound. We therefore pursue a more direct approach. Specifically, if we are able to characterize the limiting probability distribution of $\boldsymbol{w}_{k,i}$ directly, then we can evaluate the above deviation probability more precisely. To this end, we will show in this paper that the limiting distribution approaches a zero mean Gaussian random distribution with a covariance matrix that is *identical* to the one that is obtained under a *centralized* stochastic-gradient strategy. It is a useful conclusion that the distributed solution is able to recover the same error covariance matrix as the centralized solution.

Analysis of the limiting probability distribution of the error quantity for iterative algorithms can be traced back to the pioneering works [14], [15], where the authors used the moment method and the characteristic function method to derive the asymptotic probability distribution of the single-agent Robbins-Monro and Kiefer-Wolfowitz algorithms under *diminishing* step-size rules. Reference [16] provides a useful review of these algorithms and their analysis. The results in [14], [15] show that the distribution is asymptotically Gaussian. The same conclusion has been extended to distributed consensus estimation in [3], [4] also for the case of diminishing step-sizes.

Studies on the asymptotic distribution of the error quantities under *constant* step-size adaptation are largely unavailable in the literature. While [17] argued that the error vector in stand-alone LMS adaptation converges in distribution, the resulting distribution was not characterized. This question was addressed in [18], which derived an expression for the characteristic function of the limiting distribution and showed that it was *not* generally Gaussian. Reference [18] further showed that the limiting distribution can be approximated by a Gaussian for sufficiently small step-sizes. Therefore, the main challenge that arises in the constant step-size case, for both stand-alone agents and networked agents, is that the error quantities are generally *not exactly* Gaussian in steady-state. Moreover, the generalized central limit theorem developed in [14]–[16] cannot be applied directly to the constant step-size regime. Accordingly, although we are inspired by the useful

This work was supported in part by NSF grant CCF-1011918. Emails: cjs09@ucla.edu and sayed@ee.ucla.edu

results of [14]–[16], we nevertheless need to pursue a modified approach to study the steady-state probability distribution of distributed adaptation strategies under *constant* step-sizes.

## II. Distributed Stochastic Optimization

We consider a connected network of $N$ agents that are linked together through a topology. Each agent $k$ implements a distributed algorithm of the following form to update its state vector from $\boldsymbol{w}_{k,i-1}$ to $\boldsymbol{w}_{k,i}$ [10]–[12]:

$$\boldsymbol{\phi}_{k,i-1} = \sum_{l=1}^{N} a_{1,lk}\boldsymbol{w}_{l,i-1} \tag{2}$$

$$\boldsymbol{\psi}_{k,i} = \sum_{l=1}^{N} a_{0,lk}\boldsymbol{\phi}_{l,i-1} - \mu_k\hat{\boldsymbol{s}}_{k,i}(\boldsymbol{\phi}_{k,i-1}) \tag{3}$$

$$\boldsymbol{w}_{k,i} = \sum_{l=1}^{N} a_{2,lk}\boldsymbol{\psi}_{l,i} \tag{4}$$

where $\boldsymbol{w}_{k,i} \in \mathbb{R}^M$ is the state of agent $k$ at time $i$, usually an estimate for the solution of some optimization problem, $\boldsymbol{\phi}_{k,i-1} \in \mathbb{R}^M$ and $\boldsymbol{\psi}_{k,i} \in \mathbb{R}^M$ are intermediate variables generated at node $k$ before updating to $\boldsymbol{w}_{k,i}$, $\mu_k$ is a non-negative constant step-size parameter used by node $k$, and $\hat{\boldsymbol{s}}_{k,i}(\cdot)$ is an $M \times 1$ update vector function at node $k$. The combination coefficients $a_{1,lk}$, $a_{0,lk}$, and $a_{2,lk}$ in (2)–(4) are nonnegative weights that each node $k$ assigns to the information arriving from node $l$. Let $A_1$, $A_0$ and $A_2$ denote the matrices that collect $\{a_{1,lk}\}$, $\{a_{0,lk}\}$ and $\{a_{2,lk}\}$, respectively; these matrices are required to satisfy:

$$\mathbb{1}^T A_1 = \mathbb{1}^T,\ \mathbb{1}^T A_0 = \mathbb{1}^T,\ \mathbb{1}^T A_2 = \mathbb{1}^T \tag{5}$$

$$a_{1,lk} \geq 0, \qquad a_{0,lk} \geq 0, \qquad a_{2,lk} \geq 0 \tag{6}$$

$$a_{1,lk} = a_{2,lk} = a_{0,lk} = 0, \qquad \text{if } l \notin \mathcal{N}_k \tag{7}$$

Observe from (7) that the combination coefficients are zero if $l \notin \mathcal{N}_k$, where $\mathcal{N}_k$ denotes the set of neighbors of node $k$. Different choices of the combination policies will lead to different versions of distributed algorithms, such as the adapt-then-combine (ATC) strategy, combine-then-adapt (CTA) strategy, and consensus strategy — see [7] for an overview.

We argued in [12] that each estimate $\boldsymbol{w}_{k,i}$ generated by (2)–(4) converges in the mean-square-error sense to the vector $w^o$ that corresponds to the unique solution of the following algebraic equation:

$$\sum_{k=1}^{N} p_k s_k(w) = 0 \tag{8}$$

where $s_k(w)$ denotes the expected value of $\hat{\boldsymbol{s}}_{k,i}(w)$ defined by (12) further ahead, $p_k$ is the $k$th entry of the following vector:

$$p \triangleq \text{col}\{\pi_1\beta_1, \ldots, \pi_N\beta_N\} \tag{9}$$

$\pi_k$ is the $k$th entry of the vector $\pi \triangleq A_2\theta$, $\theta$ is the right eigenvector of the matrix $A = A_1A_0A_2$ that corresponds to the eigenvalue at one, $\beta_k \triangleq \mu_k/\mu_{\max}$, and $\mu_{\max} \triangleq \max \mu_k$. Furthermore, we also argued in [12] that the asymptotic covariance matrix of the error vector $\tilde{\boldsymbol{w}}_{k,i} \triangleq w^o - \boldsymbol{w}_{k,i}$ is $\mu_{\max}\cdot\Pi_0$, where the matrix $\Pi_0$ is the solution to the following Lyapunov equation:

$$H_c\Pi_0 + \Pi_0 H_c^T = (p^T \otimes I_M)\mathcal{R}_v(w^o)(p \otimes I_M) \tag{10}$$

where $\mathcal{R}_v(w^o)$ is the covariance matrix of $\hat{\boldsymbol{s}}_i(w^o)$ and

$$H_c \triangleq \sum_{k=1}^{N} p_k \nabla_{w^T} s_k(w^o) \tag{11}$$

Moreover, when $\mu_k \equiv \mu$ for all agents, the above error covariance matrix of the distributed strategy coincides with the asymptotic error covariance matrix that results from a centralized stochastic-gradient implementation using step-size $\mu' = \mu/N$ [12].

## III. Modeling Assumptions

In this section, we list the assumptions that are needed to establish the main result (Theorem 1); these conditions are of the same nature (and generally relaxations) of similar conditions often used in the analysis of the convergence behavior of distributed strategies in the literature (see, e.g., [3], [4], [11], [12], [19], [20]). For explanations on why these assumptions are justified and how they relate to assumptions used in prior studies in the literature, the readers are referred to [10]–[12].

*Assumption 1 (Strongly connected network):* The $N \times N$ matrix product $A \triangleq A_1A_0A_2$ is assumed to be a primitive left-stochastic matrix, i.e., $A^T\mathbb{1} = \mathbb{1}$ and there exists a finite integer $j_o$ such that all entries of $A^{j_o}$ are strictly positive. ■

*Assumption 2 (Update vector: Randomness):* There exists an $M \times 1$ deterministic vector function $s_k(w)$ such that for all $\boldsymbol{w} \in \mathcal{F}_{i-1}$:

$$\mathbb{E}\,\{\hat{\boldsymbol{s}}_{k,i}(\boldsymbol{w})|\mathcal{F}_{i-1}\} = s_k(\boldsymbol{w}) \tag{12}$$

for all $i, k$, where $\mathcal{F}_{i-1}$ denotes the past history of iterates $\{\boldsymbol{w}_{k,j}\}$ for $j \leq i-1$ and all $k$. Furthermore, there exist $\alpha \geq 0$ and $\sigma_v^2 \geq 0$ such that for all $i, k$ and $\boldsymbol{w} \in \mathcal{F}_{i-1}$:

$$\mathbb{E}\left\{\|\hat{\boldsymbol{s}}_{k,i}(\boldsymbol{w}) - s_k(\boldsymbol{w})\|^2\right\} \leq \alpha\cdot\mathbb{E}\|s_k(\boldsymbol{w})\|^2 + \sigma_v^2 \tag{13}$$

■

*Assumption 3 (Update vector: Lipschitz):* There exists a nonnegative $\lambda_U$ such that for all $x, y \in \mathbb{R}^M$ and all $k$:

$$\|s_k(x) - s_k(y)\| \leq \lambda_U \cdot \|x - y\| \tag{14}$$

■

*Assumption 4 (Update vector: Strong monotonicity):* Let $p_k$ denote the $k$th entry of the vector $p$ defined in (9). There exists $\lambda_L > 0$ such that for all $x, y \in \mathbb{R}^M$:

$$(x-y)^T \cdot \sum_{k=1}^{N} p_k\Big[s_k(x) - s_k(y)\Big] \geq \lambda_L \cdot \|x - y\|^2 \tag{15}$$

■

*Assumption 5 (Jacobian matrix: Lipschitz):* Assume each $s_k(w)$ is differentiable with respect to $w$ and satisfies:

$$\|\nabla_w s_k(x) - \nabla_w s_k(y)\| \leq \lambda_H \cdot \|x - y\|, \qquad \forall\ x,\ y \quad (16)$$

■

*Assumption 6 (Steady-state distribution):* Let $\boldsymbol{v}_i(w)$ denote the $MN \times 1$ global vector that collects the statistical fluctuations (gradient noise) accross all agents:

$$\boldsymbol{v}_i(w) \triangleq \mathrm{col}\{\hat{\boldsymbol{s}}_{1,i}(w) - s_1(w), \ldots \hat{\boldsymbol{s}}_{N,i}(w) - s_N(w)\} \quad (17)$$

For any $\boldsymbol{w} \in \mathcal{F}_{i-1}$, we introduce the covariance matrix:

$$\mathcal{R}_{v,i}(\boldsymbol{w}) \triangleq \mathbb{E}\left\{\boldsymbol{v}_i(\boldsymbol{w})\boldsymbol{v}_i^T(\boldsymbol{w})\big|\mathcal{F}_{i-1}\right\} \quad (18)$$

We assume that, in the limit, the second-order moment becomes invariant and tends to

$$\mathcal{R}_v \triangleq \lim_{i\to\infty} \mathcal{R}_{v,i}(w^o) \quad (19)$$

where $w^o$ denotes the limit point specified by (8). ■

The above assumptions are sufficient to characterize rather fully the convergence rate and the mean-square-error performance of the distributed strategies (2)–(4) in the constant step-size regime [12]. Next, we introduce two additional assumptions that will allow us to characterize the steady-state *probability distribution* of the error vectors. Similar assumptions were also assumed in [16, p. 147] and [15] for the study of the asympotic probability distribution of *stand-alone* stochastic approximation algorithms with diminishing step-sizes.

*Assumption 7 (Noise covariance matrix: Lipschitz):* The noise covariance matrix $\mathcal{R}_{v,i}(w)$ satisfies a Lipschitz condition, i.e., there exists a $\lambda_v \geq 0$ such that

$$\|\mathcal{R}_{v,i}(x) - \mathcal{R}_{v,i}(y)\| \leq \lambda_v \cdot \|x - y\| \quad (20)$$

for any $x, y \in \mathbb{R}^{MN}$ and all $i \geq 0$. ■

*Assumption 8 (Regularity of gradient noise):* The gradient noise process $\boldsymbol{v}_i(w)$ satisfies:

$$\lim_{\tau\to\infty} \limsup_{i\to\infty} \mathbb{E}\left[\|\boldsymbol{v}_i(w)\|^2 \cdot \mathbb{I}_{\boldsymbol{v}_i}(\tau)\right] = 0 \quad (21)$$

where $\mathbb{I}_{\boldsymbol{x}}(\tau)$ denotes the indicator function:

$$\mathbb{I}_{\boldsymbol{x}}(\tau) = 1 \text{ if } \|\boldsymbol{x}\| > \tau, \qquad \mathbb{I}_{\boldsymbol{x}}(\tau) = 0 \text{ if } \|\boldsymbol{x}\| \leq \tau \quad (22)$$

■

Assumption 7 requires the covariance matrix of the gradient noise to be smooth and Assumption 8 is a condition on the tail distribution of the gradient noise. Specifically, it requires that, in steady-state, the tail of the probability distribution of the gradient noise decays at a sufficiently fast speed.

## IV. Summary of Main Result

Let $\boldsymbol{w}_i = \mathrm{col}\{\boldsymbol{w}_{1,i}, \ldots, \boldsymbol{w}_{N,i}\}$ collect the iterates across the network at time $i$ and introduce $\tilde{\boldsymbol{w}}_i \triangleq \mathbb{1} \otimes w^o - \boldsymbol{w}_i$ to represent the network error vector. Theorem 1 and Eq. (33) further ahead will establish that, for sufficiently small step-sizes, the steady-state normalized error vector, $\tilde{\boldsymbol{w}}_i/\sqrt{\mu_{\max}}$, is close to the following Gaussian distribution:

$$\frac{\tilde{\boldsymbol{w}}_i}{\sqrt{\mu_{\max}}} \sim \mathcal{N}(0,\ \mathbb{1}\mathbb{1}^T \otimes \Pi_0), \qquad i \gg 1 \quad (23)$$

where $\Pi_0$ is the solution to the Lyapunov equation (10). It then follows that the asymptotic estimator $\boldsymbol{w}_i$ is approximately distributed according to

$$\boldsymbol{w}_i \sim \mathcal{N}(\mathbb{1} \otimes w^o,\ \mu_{\max} \cdot \mathbb{1}\mathbb{1}^T \otimes \Pi_0), \quad i \gg 1 \quad (24)$$

Likewise, the asymptotic marginal distribution for the steady-state iterates at each agent $k$ will be given by

$$\boldsymbol{w}_{k,i} \sim \mathcal{N}(w^o,\ \mu_{\max} \cdot \Pi_0), \qquad i \gg 1 \quad (25)$$

The above results imply that the estimate at each agent fluctuates around the optimal solution $w^o$ according to a Gaussian distribution, where the covariance matrix is the same value as that of a centralized stochastic-gradient strategy. The result holds regardless of the specific topology of the network.

In practice, we are often interested in the probability that the iterate at each agent deviates from the optimal solution $w^o$ by a certain distance. A direct consequence of the results derived in this paper is that we can now evaluate such deviations by relying on the above distributions and their mean and variance parameters. For example, for a given precision level $\epsilon > 0$ and a positive semi-definite weighting matrix $\Sigma$, we have

$$\begin{aligned}\Pr\left\{\|\boldsymbol{w}_{k,i} - w^o\|_\Sigma > \epsilon\right\} &= \Pr\left\{\|\tilde{\boldsymbol{w}}_{k,i}\|_\Sigma^2 > \epsilon^2\right\} \\ &\approx \int_{\|x\|_\Sigma^2 \geq \epsilon^2} p_{\tilde{\boldsymbol{w}}}(x)dx \end{aligned} \quad (26)$$

where $p_{\tilde{\boldsymbol{w}}}$ denotes the asymptotic Gaussian distribution of the iterates $\{\tilde{\boldsymbol{w}}_{k,i}\}$ according to (25). Calculations of the form (26) are useful in many contexts. For example, they can be used to determine the decision thresholds in distributed detection problems [21]. The distribution of $\boldsymbol{w}_{k,i}$ enables the computation of the decision thresholds as a function of the false alarm rate in closed-form.

## V. Arguments and Analysis

Due to space limitations, we are only able to highlight the main steps in the argument; proofs are omitted for brevity. To begin with, we know from (10) and [12] that the mean-square-error $\mathbb{E}\|\tilde{\boldsymbol{w}}_i\|^2$ is on the order of $O(\mu_{\max})$, which will be small for small $\mu_{\max}$. We therefore examine the probability distribution of the normalized error vector $\tilde{\boldsymbol{w}}_i/\sqrt{\mu_{\max}}$ as $i \to \infty$. We pursue this task in two steps:

- First, we show that $\tilde{\boldsymbol{w}}_i/\sqrt{\mu_{\max}}$ is close to $\tilde{\boldsymbol{w}}_{s,i}/\sqrt{\mu_{\max}}$ in high probability, where $\tilde{\boldsymbol{w}}_{s,i}$ is defined by (27) below. We also argue that the cumulative distribution functions

(CDFs) of these two random vectors are close to each other.

- Then, we show that the distribution of $\check{\boldsymbol{w}}_{s,i}/\sqrt{\mu_{\max}}$ is close to a zero mean Gaussian random vector with the covariance matrix being $\Pi_0$. This is proved by showing that the characteristic functions of these two distributions are close to each other.

*Lemma 1 (Close in probability):* Let $\check{\boldsymbol{w}}_{s,i}$ denote the following $M \times 1$ vector

$$\check{\boldsymbol{w}}_{s,i} = \mu_{\max} \sum_{n=0}^{i-1} B_c^n (p^T \otimes I_M) \boldsymbol{v}_{i-n} \tag{27}$$

where

$$B_c \triangleq I - \mu_{\max} H_c \tag{28}$$

$$\boldsymbol{v}_i \triangleq \boldsymbol{v}_i(\boldsymbol{\phi}_i) \tag{29}$$

$$\boldsymbol{\phi}_i \triangleq \mathrm{col}\{\boldsymbol{\phi}_{1,i}, \ldots \boldsymbol{\phi}_{N,i}\} \tag{30}$$

Then, for any $\epsilon > 0$, it holds that

$$\limsup_{i\to\infty} \Pr\left\{ \left\| \frac{\tilde{\boldsymbol{w}}_i}{\sqrt{\mu_{\max}}} - \mathbb{1} \otimes \frac{\check{\boldsymbol{w}}_{s,i}}{\sqrt{\mu_{\max}}} \right\| > \epsilon \right\} \leq O\left( \frac{\sqrt{\mu_{\max}}}{\epsilon} \right) + O\left( \frac{\mu_{\max}}{\epsilon^2} \right) \tag{31}$$

*Proof:* Omitted for brevity. ∎

An important observation that follows from the above lemma is that, if a random variable $\boldsymbol{x}$ is close to another random variable $\boldsymbol{y}$ in high probability, as indicated by (31), then their probability distributions are close to each other as well. This is analogous to the fact that "convergence in probability implies convergence in distribution". This statement can be made rigorous but the argument is omitted.

Next, we call upon the following lemma to show that the distribution of $\check{\boldsymbol{w}}_{s,i}/\sqrt{\mu_{\max}}$ can be arbitrarily close to a Gaussian as $i \to \infty$.

*Lemma 2 (Asymptotic Gaussian distribution):* Suppose $\boldsymbol{g}$ is a zero mean Gaussian random vector with covariance matrix $\Pi_0$. Then, for any given $\nu$, we have

$$\lim_{\mu_{\max}\to 0} \limsup_{i\to\infty} \left| F_{\mathbb{1}\otimes \frac{\check{\boldsymbol{w}}_{s,i}}{\sqrt{\mu_{\max}}}}(\nu) - F_{\mathbb{1}\otimes \boldsymbol{g}}(\nu) \right| = 0 \tag{32}$$

where the notation $F_{\boldsymbol{x}}(\nu)$ denotes the cumulative distribution of the random variable $\boldsymbol{x}$, i.e., $\Pr(\boldsymbol{x} \preceq \nu)$.

*Proof:* Omitted for brevity. ∎

Lemmas 1 and 2 can be combined to establish the main theorem.

*Theorem 1 (Asymptotic probability distribution):* Let $\boldsymbol{g}$ be the same Gaussian random vector defined in Lemma 2. Then the CDF of the normalized error vector $\tilde{\boldsymbol{w}}_i/\sqrt{\mu_{\max}}$ satisfies

$$\lim_{\mu_{\max}\to 0} \limsup_{i\to\infty} \left| F_{\frac{\tilde{\boldsymbol{w}}_i}{\sqrt{\mu_{\max}}}}(\nu) - F_{\mathbb{1}\otimes \boldsymbol{g}}(\nu) \right| = 0 \tag{33}$$

*Proof:* Omitted for brevity. ∎

In the above theorem, it is important to note that both $i \to \infty$ and $\mu_{\max} \to 0$. If we do not have sufficiently small step-sizes, then the asymptotic distribution of $\tilde{\boldsymbol{w}}_i/\sqrt{\mu_{\max}}$ is not necessarily Gaussian. For example, in the case of a stand-alone agent running the LMS recursion, it was shown in [18, Eq. (23)] that the asymptotic characteristic function of $\tilde{\boldsymbol{w}}_i/\sqrt{\mu_{\max}}$ has a form that is different than that of a Gaussian distribution! However, when the $\mu_{\max}$ becomes small, the characteristic function derived in [18] was shown to approach that of a Gaussian distribution.

## REFERENCES


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

[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. Kar, J. M. F. Moura, and K. Ramanan, "Distributed parameter estimation in sensor networks: Nonlinear observation models and imperfect communication," *IEEE Trans. Inf. Theory*, vol. 58, no. 6, pp. 3575–3605, June 2012.

[5] 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.

[6] 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.

[7] A. H. Sayed, "Diffusion adaptation over networks," *in* E-Reference Signal Processing, R. Chellapa and S. Theodoridis, *editors,* Elsevier, 2013 [Also available online as arXiv:1205.4220v1, May 2012].

[8] 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.

[9] 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.

[10] 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.

[11] ——, "Distributed Pareto optimization via diffusion adaptation," *IEEE J. Sel. Topics Signal Process.*, vol. 7, no. 2, pp. 205–220, Apr. 2013.

[12] ——, "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.

[13] V. H. Nascimento and A. H. Sayed, "On the learning mechanism of adaptive filters," *IEEE Trans. Signal Process.*, vol. 48, no. 6, pp. 1609–1625, Jun. 2000.

[14] K. L. Chung, "On a stochastic approximation method," *Ann. Math. Stat.*, vol. 25, no. 3, pp. 463–483, 1954.

[15] J. Sacks, "Asymptotic distribution of stochastic approximation procedures," *The Annals of Mathematical Statistics*, vol. 29, no. 2, pp. 373–405, Jun. 1958.

[16] M. B. Nevelson and R. Z. Hasminskii, *Stochastic Approximation and Recursive Estimation*. American Mathematical Society, 1976.

[17] R. Bitmead, "Convergence in distribution of lms-type adaptive parameter estimates," *IEEE Trans. Autom. Control*, vol. 28, no. 1, pp. 54–60, Jan. 1983.

[18] X. Zhao and A. H. Sayed, "Probability distribution of steady-state errors and adaptation over networks," in *Proc. IEEE Statistical Signal Processing Workshop (SSP)*, Nice, France, Jun. 2011, pp. 253–256.

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

[20] B. T. Polyak and Y. Z. Tsypkin, "Pseudogradient adaptation and training algorithms," *Automation and Remote Control*, vol. 12, pp. 83–94, 1973.

[21] F. S. Cattivelli and A. H. Sayed, "Distributed detection over adaptive networks using diffusion adaptation," *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 1917–1932, May 2011.