CONVERGENCE OF A DISTRIBUTED PARAMETER ESTIMATOR FOR SENSOR NETWORKS WITH LOCAL AVERAGING OF THE ESTIMATES

P. Bianchi, G. Fort, W. Hachem, J. Jakubowicz

LTCI, TELECOM ParisTech / CNRS, 46 rue Barrault 75634 Paris Cedex 13, France.
email: name@telecom-paristech.fr

ABSTRACT

The paper addresses the convergence of a decentralized Robbins-Monro algorithm proposed by [7] for networks of agents. This algorithm combines local stochastic approximation steps for finding the root of an objective function, and a gossip step for consensus seeking between agents. We provide verifiable sufficient conditions on the stochastic approximation procedure and on the network so that the decentralized Robbins-Monro algorithm converges to a consensus. We also prove that the limit points of the algorithm correspond to the roots of the objective function. We apply our results to Maximum Likelihood estimation in sensor networks.

1. INTRODUCTION

In many applications, one is interested in finding the roots of a given equation $h(\theta) = 0$. A traditional approach is to solve this equation iteratively (see *e.g.* [6] for deterministic procedures and [5] for stochastic ones). When h is unknown in closed form but only a noisy stochastic version $H(\theta, X)$ is available, the Robbins-Monro algorithm finds roots of h by a stochastic approximation procedure [8]. In signal processing, a typical application is given by Maximum Likelihood (ML) estimators based on a gradient search. In this particular case, the aim is to search for the stationary points of the Kullback-Leibler divergence between the true distribution of the observations and a distribution in a parametric family. In this example, $H(\theta, X)$ is the gradient w.r.t. θ of the log-likelihood function associated with a given observation X.

In this paper, we investigate decentralized Robbins-Monro algorithms, as introduced in [7, 11]. Consider a network composed of N nodes. At time n, each node $i \in \{1, \dots, N\}$ has its own local iterate $\theta_{n,i}$. Each node observes a local random variable $X_{n+1,i}$ and uses this observation to update its local iterate based on a Robbins-Monro dynamic. In addition, nodes are able to communicate with some neighbours at certain (possibly random) moments. When such a communication occurs, the agents find an agreement on the value of the iterate, and substitute the value of their current iterate with this agreement. We thus address the convergence of an algorithm which combines a stochastic approximation step (see e.g. [5]) and a gossip step (see e.g. [1, 3]). Distributed consensus have been the object of numerous works in the recent signal processing literature [10]. Gossip algorithms for consensus seeking are a class of iterative methods which compute the average of a given deterministic vector in a distributed fashion. More recently, several authors manage to cast the classical stochastic approximation approach for finding roots of an objective function into the framework of distributed consensus [4, 7, 9, 11].

The main contribution of this paper is to provide a complete convergence analysis of decentralized Robbins-Monro algorithms. This paper generalizes earlier works of [7]. First, our result applies when the mean field which governs the stochastic approximation step is not bounded. Second, it does not require the introduction of a projection step which is known to modify the set of the limit points of the algorithm. Finally, we provide sufficient conditions in the noise function H which are fully explicit and easily verifiable.

The paper is organized as follows. Section 2 is devoted to the description of the algorithm and to the statement of our assumptions. Convergence results are stated in Section 3. Section 4 provides an application of our results to ML estimation in a distributed sensor network. Numerical results are given in Section 5.

2. ALGORITHM DESCRIPTION AND ASSUMPTIONS

2.1. Algorithm

Consider a network composed by $N \geq 1$ nodes, and assume that node $i \in \{1, \dots, N\}$ observes the random variable $X_{n,i}$ at time n. Each node i generates a stochastic process $(\theta_{n,i})_{n \geq 1}$ in $\mathbb{R}^d$ using a two-steps iterative algorithm:

`[Local step]` Node i generates at time n a temporary iterate $\tilde{\theta}_{n,i}$ given by

$$\tilde{\theta}_{n,i} = \theta_{n-1,i} + \gamma_n \, H_i(\theta_{n-1,i} \, ; X_{n,i}) \, , \qquad (1)$$

where γ_n is a deterministic positive step size and $H_i(\theta_{n-1,i} \, ; X_{n,i})$ is some increment chosen as a function of

This work is partially supported by the French National Research Agency, under the program ANR-07 ROBO 0002

the previous iterate and the current observation.
`[Gossip step]` Node i is able to observe the values $\tilde{\theta}_{n,j}$ of some other j's and computes the weighted average:

$$\theta_{n,i} = \sum_{j=1}^{N} w_n(i,j)\,\tilde{\theta}_{n,j}$$

where $W_n := [w_n(i,j)]_{i,j=1}^N$ is a stochastic matrix.

It is convenient to cast this algorithm into a vector form. Assume that for any $n \geq 1$, $X_{n,i} \in \mathbb{R}^{m_i}$. Define the function $H : \mathbb{R}^{dN} \times \mathbb{R}^{\sum m_i} \to \mathbb{R}^{dN}$ as

$$H(\boldsymbol{\theta}; x) := \left(H_1(\theta_1; x_1)^T, \cdots, H_N(\theta_N; x_N)^T\right)^T \ .$$

where ^{T} denotes transposition, $x = (x_1^T, \ldots, x_N^T)^T$ and $\boldsymbol{\theta} = (\theta_1^T, \ldots, \theta_N^T)^T$. Define the random vectors $\boldsymbol{\theta}_n$ and X_n as $\boldsymbol{\theta}_n := (\theta_{n,1}^T, \ldots, \theta_{n,N}^T)^T$ and $X_n = (X_{n,1}^T, \ldots, X_{n,N}^T)^T$. The algorithm reduces to:

$$\boldsymbol{\theta}_n = (W_n \otimes I_d)(\boldsymbol{\theta}_{n-1} + \gamma_n H(\boldsymbol{\theta}_{n-1}; X_n)) \ , \qquad (2)$$

where $\otimes$ denotes the Kroenecker product and I_d is the $d \times d$ identity matrix.

2.2. Model Assumptions

The time varying communication network between the nodes is represented by the sequence of random matrices $(W_n)_{n\geq 1}$. Denote by $\mathbb{1}_N$ the $N \times 1$ vector whose components are all equal to one. It is assumed that:

A1 a) Matrix W_n is doubly stochastic: $W_n \mathbb{1} = W_n^T \mathbb{1} = \mathbb{1}$.
b) Matrices $(W_n)_{n\geq 1}$ are i.i.d. and the spectral radius of $\mathbb{E}(W_1) - \mathbb{1}\mathbb{1}^T/N$ is strictly less than one.
c) For any functions f, g,

$$\mathbb{E}[f(W_{n+1})g(X_{n+1})|\boldsymbol{\theta}_0, X_{1:n}, W_{1:n}] = \mathbb{E}[f(W_1)]\,\mathbb{E}[g(X_{n+1})|\boldsymbol{\theta}_n] \qquad (3)$$

Condition **A1**a) is satisfied provided that the nodes coordinate their weights. Coordination schemes are discussed in [3, 7]. The condition also holds in case of asynchronous networks (see [1, 3] for details and see Section 4 for a brief discussion). Condition **A1**b) can be interpreted as follows. The intuitive idea behind gossip algorithms is that $\mathbb{E}(W_n)$ should be close enough to the projector $\mathbb{1}\mathbb{1}^T/N$ on the line $\{t\mathbb{1} : t \in \mathbb{R}\}$ so that the algorithm (2) reaches an average consensus. Condition **A1**b) on the spectral radius ensures that the amount of information exchanged in the network remains sufficient in order to reach a consensus. The hypothesis that matrices W_n are identically distributed can be weakened in order to cover the case where the average number of communications between nodes is likely to vary in time and, possibly, to vanish as n increases. In that case, the condition on the spectral radius must be somewhat reinforced (see [2]). Condition **A1**c) implies that r.v. W_{n+1} and X_{n+1} are independent conditionally to the past. In addition, the conditional distribution of X_{n+1} depends on the past only through $\boldsymbol{\theta}_n$.

Hereafter, we use notation $\mathbb{E}_{\boldsymbol{\theta}_n}$ for $\mathbb{E}[\,.\,|\boldsymbol{\theta}_n]$.

3. CONVERGENCE ANALYSIS

The convergence analysis relies on the existence of a Lyapunov function V for the function h i.e. a function such that $\nabla V^T\, h \leq 0$. In this framework, the classical approach to prove the convergence of a stochastic approximation procedure to the roots of h is to prove *(a)* that, with probability one (w.p.1), the path remains in a compact set and *(b)* that the sequence converges to the set $\mathcal{L} := \{\nabla V^T\, h = 0\}$. To that goal, regularity conditions on the functions H and V, and on the set of the limit points $\mathcal{L}$ are required.

3.1. Notations

We denote by $J := (\mathbb{1}\mathbb{1}^T/N) \otimes I_d$ the projector onto the consensus subspace $\left\{\mathbb{1} \otimes \theta : \theta \in \mathbb{R}^d\right\}$ and by $J^\perp := I_{dN} - J$ the projector onto the orthogonal subspace. For any vector $\boldsymbol{\theta} \in \mathbb{R}^{dN}$, remark that $\boldsymbol{\theta} = \mathbb{1} \otimes \langle\boldsymbol{\theta}\rangle + J^\perp\boldsymbol{\theta}$ where

$$\langle\boldsymbol{\theta}\rangle := \frac{1}{N}(\mathbb{1}^T \otimes I_d)\boldsymbol{\theta} \qquad (4)$$

is a vector of $\mathbb{R}^d$. Equation (4) simply means that $\langle\boldsymbol{\theta}\rangle = (\theta_1 + \cdots + \theta_N)/N$ in case we write $\boldsymbol{\theta} = (\theta_1^T, \ldots, \theta_N^T)^T$ for some $\theta_1, \ldots, \theta_N$ in $\mathbb{R}^d$. We introduce the *mean field* of the decentralized Robbins-Monro algorithm as the function $h : \mathbb{R}^d \to \mathbb{R}^d$ given by:

$$h(\theta) := \mathbb{E}_{\mathbb{1}\otimes\theta}\left[\langle H(\mathbb{1} \otimes \theta; X)\rangle\right] \ , \qquad (5)$$

where we recall that $\langle H(\boldsymbol{\theta}; x)\rangle = \frac{1}{N}(\mathbb{1}^T \otimes I_d)H(\boldsymbol{\theta}; x)$ is the average of $H(\boldsymbol{\theta}; x)$ (see Eq.(4)).

3.2. Convergence result

Denote by $|x|$ the Euclidean norm of a vector $x \in \mathbb{R}^l$, and by ∇ the gradient operator. It is assumed that

A2 The deterministic sequence $(\gamma_n)_{n\geq 1}$ is positive and such that $\sum_n \gamma_n^2 < \infty$, $\sum_n \gamma_n = \infty$.

A3 There exists a function $V : \mathbb{R}^d \to \mathbb{R}^+$ such that:
a) V is differentiable and ∇V is a Lipschitz function.
b) For any $\theta \in \mathbb{R}^d$, $\nabla V(\theta)^T h(\theta) \leq 0$.
c) There exists a constant C_1, such that for any $\theta \in \mathbb{R}^d$, $|\nabla V(\theta)|^2 \leq C_1(1 + V(\theta))$.
d) For any $M > 0$, the level set $\{\theta \in \mathbb{R}^d : V(\theta) \leq M\}$ is compact.
e) The set $\mathcal{L} := \{\theta \in \mathbb{R}^d : \nabla V(\theta)^T h(\theta) = 0\}$ is bounded.
f) $V(\mathcal{L})$ has an empty interior.

Assumption **A2** is classical in stochastic approximation and is satisfied for example with $\gamma_n \propto n^{-a}$ for $a \in (1/2, 1]$. Assumption **A3**b) means that V is a Lyapunov function for the mean field h. When h is known (and continuous), **A3** combined with the condition $\sum_n \gamma_n = +\infty$ allows to prove the convergence of the deterministic sequence $t_{n+1} = t_n + \gamma_{n+1} h(t_n)$ to the set $\mathcal{L}$. When h is unknown and replaced by a stochastic approximation H, the limiting behavior of the noisy algorithm is the same provided H satisfies some regularity conditions and the step-size sequence satisfies $\sum_n \gamma_n^2 < \infty$. We assume:

A4 a) There exists a constant C_2 such that for any $\boldsymbol{\theta} \in \mathbb{R}^{dN}$,

$$\mathbb{E}_{\boldsymbol{\theta}}\left[|H(\boldsymbol{\theta}; X)|^2\right] \leq C_2\left(1 + V(\langle\boldsymbol{\theta}\rangle) + |J^{\perp}\boldsymbol{\theta}|^2\right)$$
$$\mathbb{E}_{\boldsymbol{\theta}}\left|\langle H(\boldsymbol{\theta}; X)\rangle - \langle H(J\boldsymbol{\theta}; X)\rangle\right| \leq C_3 |J^{\perp}\boldsymbol{\theta}|$$
$$\left|\mathbb{E}_{\boldsymbol{\theta}}\langle H(\boldsymbol{\theta}; X)\rangle - \mathbb{E}_{J\boldsymbol{\theta}}\langle H(J\boldsymbol{\theta}; X)\rangle\right| \leq C_4 |J^{\perp}\boldsymbol{\theta}| \; .$$

b) Function h is continuous on $\mathbb{R}^d$.

Under **A1**, **A3**a-c) and **A4**, we prove that the sequence $(\boldsymbol{\theta}_n - \mathbb{1} \otimes \langle\boldsymbol{\theta}_n\rangle)_{n\geq 1}$ converges almost-surely (and in L^2) to zero, and the sequence $(\langle\boldsymbol{\theta}_n\rangle)_{n\geq 1}$ enters infinitely often some level set $\{V \leq M\}$. Conditions **A3**b-e), **A4** and **A2** imply that, almost-surely, *(a)* the sequence $(\langle\boldsymbol{\theta}_n\rangle)_{n\geq 1}$ remains in a neighborhood of $\mathcal{L}$ thus implying that the sequence remains in a compact set of $\mathbb{R}^d$ and *(b)* the sequence $(V(\langle\boldsymbol{\theta}_n\rangle))_{n\geq 1}$ converges to a connected component of $V(\mathcal{L})$. Finally, **A3**f) implies the convergence of $(\langle\boldsymbol{\theta}_n\rangle)_{n\geq 1}$ to a connected component of $V(\mathcal{L})$. The proof of Theorem 1 is omitted due to lack of space and will be provided in an extended version of this paper (see [2]).

Define the distance $\mathsf{d}(\theta, A)$ between a point $\theta \in \mathbb{R}^d$ and a subset $A \subset \mathbb{R}^d$ by $\mathsf{d}(\theta, A) = \inf\{|\theta - \varphi| \; : \varphi \in A\}$.

Theorem 1 *Assume* **A1**, **A2**, **A3** *and* **A4** *and consider the algorithm (2). Then, w.p.*1*,*

$$\lim_{n\to\infty} |\boldsymbol{\theta}_n - \mathbb{1} \otimes \langle\boldsymbol{\theta}_n\rangle| = 0 \; , \qquad \lim_{n\to\infty} \mathsf{d}(\langle\boldsymbol{\theta}_n\rangle, \mathcal{L}) = 0 \; .$$

*Moreover, w.p.*1*,* $(\langle\boldsymbol{\theta}_n\rangle)_{n\geq 1}$ *converges to a connected component of* $\mathcal{L}$*.*

Theorem 1 states that, almost surely, the vector of iterates $\boldsymbol{\theta}_n$ converges to the consensus space as $n \to \infty$. Moreover, the average iterate of the network converge to some connected component of $\mathcal{L}$. When $\mathcal{L}$ is finite, Theorem 1 implies that, almost surely, $\langle\boldsymbol{\theta}_n\rangle$ converges to some point in $\mathcal{L}$.

4. APPLICATION TO MAXIMUM LIKELIHOOD ESTIMATION

We assume that the local observations $X_{n,i} \in \mathbb{R}^{m_i}$ $(i = 1, \ldots, N)$ are block-components of $X_n \in \mathbb{R}^{\sum m_i}$. Furthermore, process $(X_n)_{n\geq 1}$ is i.i.d. with unknown p.d.f. $f^{\star}$. The aim is to use the previous algorithm in order to fit $f^{\star}$ with a probability distribution $f(\,.\,;\theta)$ chosen among a parametric family indexed by $\theta \in \mathbb{R}^d$ of the form $\prod_{i=1}^N f_i(x_i\,;\theta)$ where $x = (x_1^T, \ldots, x_N^T)^T$. To that end, we use a decentralized stochastic gradient maximum likelihood approach: we define for each i, $H_i(\theta; X_{n,i}) := \nabla_\theta \log f_i(X_{n,i}\,;\theta)$ so that the mean field h is given by $h(\theta) = (1/N)\sum_i \mathbb{E}[\nabla_\theta \log f_i(X_{n,i}\,;\theta)]$. By Theorem 1, the algorithm (2) searches for the roots of h. These roots are the stationary points of the Kullback-Leibler (KL) divergence:

$$V(\theta) := \int f^{\star}(x) \log \frac{f^{\star}(x)}{f(x;\theta)} dx \; . \tag{6}$$

Under regularity conditions on the densities f_i, $h = -(1/N)\nabla V$ so that V is a natural Lyapunov function. In this situation, the set $\mathcal{L}$ is equal to the set of stationary points of the KL divergence $\mathcal{L} = \{\theta \in \mathbb{R}^d : \nabla V(\theta) = 0\}$. Moreover, by Sard's theorem, $V(\mathcal{L})$ has an empty interior as soon as V is d times continuously differentiable.

Examples of densities f_i such that H and V satisfy **A3** and **A4** are given in the next section. By direct application of Theorem 1, sequence $(\boldsymbol{\theta}_n)_{n\geq 1}$ converges to the consensus subspace and the average estimate sequence $(\langle\boldsymbol{\theta}_n\rangle)_{n\geq 1}$ converges to $\mathcal{L}$. In particular, the decentralized ML estimator and the centralized one have the same limit points.

Comments on the Network Model. The network is described by a nondirected graph $(\mathcal{V}, \mathcal{E})$ whose vertices $\mathcal{V}$ correspond to the nodes $\{1 \ldots N\}$ and whose edges are formed by the pairs of nodes $\{i, j\}$ which are likely to communicate. Consider for instance the framework of asynchronous communications, which matches to the distributed nature of sensor networks. An example of asynchronous network model for matrices $(W_n)_{n\geq 1}$ can be found for instance in [3]. This model can be described as follows. At each time n, assume that one node i wakes up and initiates a bidirectional communication with one of its neighbour j. This event occurs with probability P_{ij}, where $P_{ij} > 0$ if and only if i and j are connected. Nodes i and j replace their local temporary estimates $\tilde{\theta}_{n,i}$ and $\tilde{\theta}_{n,j}$ respectively with the average of these two values. As a consequence $W_n = I_N - \frac{1}{2}(c_i - c_j)(c_i - c_j)^T$ where c_i denotes the ith column-vector of the canonical basis on $\mathbb{R}^N$ (this matrix has all its diagonal coefficients equal to 1 and all its nondiagonal coefficients equal to zero, except for the coefficients (i,i), (i,j), (j,i), (j,j) which are equal to $1/2$). It is straightforward to prove that Assumptions **A1**a-b) are satisfied under the above network model as soon as the graph $(\mathcal{V}, \mathcal{E})$ is connected. More involved network models are developed in [2].

5. NUMERICAL RESULTS

Consider a network formed by N fixed sensors in the unit square $[0, 1]^2$. Assume that the aim of the network is to es-

timate the geographic coordinates of D sources in $\mathbb{R}^2$. Denote by $\theta^\star$ the $D \times 1$ complex valued vector which contains the complex locations of the D sources. At each iteration n, each sensor $i = 1 \dots N$ observes a noisy version $Y_{n,i}$ of $\theta^\star$. We assume that $Y_{n,i} \sim \mathcal{CN}(\theta^\star, \text{Diag}\left(\sigma^2_{i,1} \dots \sigma^2_{i,D}\right))$ where variances $\sigma^2_{i,1} \dots \sigma^2_{i,D}$ are assumed to be perfectly known at node i (and at node i only). In our simulations, we generate $\sigma^2_{i,k}$ for each source $k = 1 \dots D$ located in $\theta^\star_k \in \mathbb{C}$ and each node i located in $z_i \in \mathbb{C}$, as $\sigma^2_{i,k} = \alpha|\theta^\star_k - z_i|^s + \delta_0$ where $s > 0$ represents a path loss exponent, α is a constant, and δ_0 is a fixed error variance. The network follows a random geographical graph model. Both sources and nodes locations are drawn independently according to the uniform distribution in the unit square. Two nodes i and j are connected iff $|i - j| < r$ for some radius r. We set $N = 20$, $D = 4$, $r = 0.2$, $s = 2$, $\alpha = 10$, $\delta_0 = 0.1$. The step size is chosen as $\gamma_n = 0.1/n$. Figure 1 provides a realization of the simulation scenario described above. Locations of sources and nodes are represented in the unit square. At

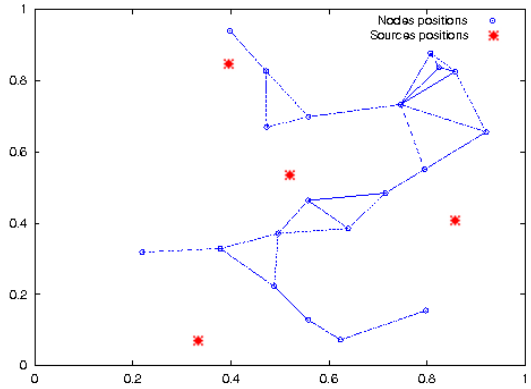

Fig. 1. One realization of the network graph and sources' locations.

iteration n, denote by $\theta_{k,i}(n)$ the complex estimate of the kth source position at the ith node. Define the average estimate of the kth source position as $\bar{\theta}_k(n) = \frac{1}{N}\sum_{i=1}^N \theta_{k,i}(n)$. Define the disagreement between nodes on the kth source as $\Delta_k(n) = (\frac{1}{N}\sum_{i=1}^N |\theta_{k,i}(n) - \bar{\theta}_k(n)|^2)^{1/2}$. Finally, define the average error as $\epsilon_k(n) = |\bar{\theta}_k(n) - \theta^\star_k|$. Figure 2 and 3 respectively represent the disagreement and the average error as a function of the number of iterations. As expected, both error converge to zero as n tends to infinity.

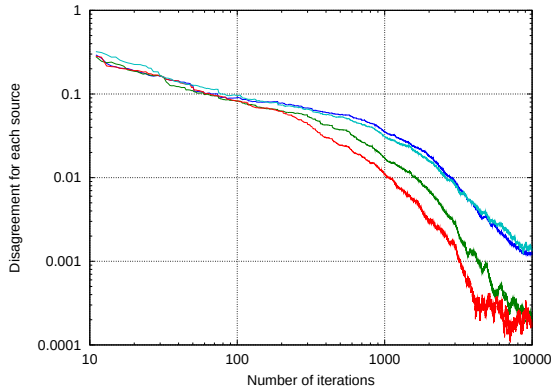

Fig. 2. Disagreement $\Delta_k(n)$ as a function of the number n of iterations $(k = 1 \dots 4)$.

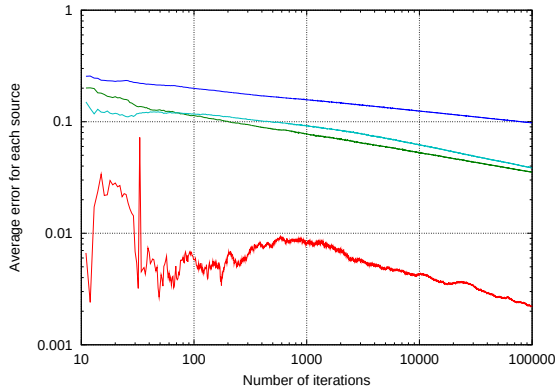

Fig. 3. Average error $\epsilon_k(n)$ as a function of the number n of iterations $(k = 1 \dots 4)$.

6. REFERENCES

[1] T.C. Aysal, M.E. Yildiz, A.D. Sarwate, and A. Scaglione. Broadcast gossip algorithms for consensus. *IEEE Trans. on Signal Processing*, 57(7):2748–2761, 2009.

[2] P. Bianchi, G. Fort, W. Hachem, J. Jakubowicz, and E. Moulines. On the convergence of decentralized robbins-monro algorithms. 2010. Work in progress.

[3] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah. Randomized gossip algorithms. *IEEE Trans. on Inform. Theory*, 52(6):2508–2530, 2006.

[4] S. Kar and J.M.F. Moura. Distributed consensus algorithms in sensor networks: Quantized data and random link failures. *IEEE Trans. on Signal Processing*, 58(3):1383–1400, 2010.

[5] H.J. Kushner and G.G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, second edition edition, 2003.

[6] D.G. Luenberger. *Linear and Nonlinear Programming*. Addison-Wesley, second edition edition, 1984.

[7] S.S. Ram, A. Nedic, and V.V. Veeravalli. Distributed stochastic subgradient projection algorithms for convex optimization. *Arxiv preprint arXiv:0811.2595*, 2008.

[8] H. Robbins and S. Monro. A stochastic approximation method. *Ann. of Mathem. Statist.*, 22(3):400–407, 1951.

[9] I.D. Schizas, G. Mateos, and G.B. Giannakis. Distributed lms for consensus-based in-network adaptive processing. *IEEE Trans. on Signal Processing*, 56(6):2365–2382, 2009.

[10] I.D. Schizas, A. Ribeiro, and G.B. Giannakis. Consensus in ad hoc WSNs with noisy links-Part I: Distributed estimation of deterministic signals. *IEEE Trans. on Signal Processing*, 56(1):350–364, 2008.

[11] S.S. Stankovic, M.S. Stankovic, and D.M. Stipanovic. Decentralized parameter estimation by consensus based stochastic approximation. In *2007 46th IEEE Conference on Decision and Control*, pages 1535–1540, 2008.