Distributed Stochastic Approximation for Constrained and Unconstrained Optimization*

Pascal Bianchi Jérémie Jakubowicz

Telecom ParisTech / CNRS-LTCI
46, rue Barrault - 75634 Paris Cedex - France
{bianchi,jakubowi}@telecom-paristech.fr

ABSTRACT

In this paper, we analyze the convergence of a distributed Robbins-Monro algorithm for both constrained and unconstrained optimization in multi-agent systems. The algorithm searches local minima of a (nonconvex) objective function which is supposed to coincide with a sum of local utility functions of the agents. The algorithm under study consists of two steps: a local stochastic gradient descent at each agent and a gossip step that drives the network of agents to a consensus. It is proved that *i)* an agreement is achieved between agents on the value of the estimate, *ii)* the algorithm converges to the set of Kuhn-Tucker points of the optimization problem. The proof relies on recent results about differential inclusions. In the context of unconstrained optimization, intelligible sufficient conditions are provided in order to ensure the stability of the algorithm. In the latter case, we also provide a central limit theorem which governs the asymptotic fluctuations of the estimate. We illustrate our results in the case of distributed power allocation for ad-hoc wireless networks.

1. INTRODUCTION

The Robbins-Monro (R-M) algorithm [1] is a widely used procedure for finding the roots of an unknown function. Its applications range from Statistics (e.g. [2]), Machine Learning (e.g. [3]), Electrical Engineering (e.g. [4]) and Communication Networks. Consider the problem of minimizing a given differentiable function f. Formally, a R-M algorithm for that sake can be summarized as an iterative scheme of the form $\theta_{n+1} = \theta_n + \gamma_{n+1}(-\nabla f(\theta_n) + \xi_{n+1})$ where the sequence $(\theta_n)_{n\in\mathbb{N}}$ will eventually converge to a local minimum of f, and where ξ_{n+1} represents a random perturbation.

In this paper, we investigate a *distributed* version of the R-M algorithm. Distributed algorithms have aroused deep interest in the fields of communications, signal processing, control, robotics, computer technology, among others. The success of distributed algorithms lies in their scalability but are often harder to analyze than their centralized counterparts. We analyze the behavior of a network of agents, represented as a graph, where each node/agent runs its own local R-M algorithm and then randomly communicates with one of its neighbors in the hope of gradually reaching a consensus over the whole network. One well-established device for reaching a consensus in a network is to use gossip algorithms [5]. Since the seminal paper of [6], *random* gossip algorithms have been widely studied as they encompass asynchronous networks with random switching graph topologies. In [5], the Authors introduce an iterative algorithm for the optimization of an objective function in a parallel setting. The method consists in an iterative gradient search combined with a gossip step. More recently, this algorithm has been studied by [7, 8] in the case where the objective function is the aggregate of some local utility functions of the agents, assuming that a given agent is only able to evaluate a (noisy version of) the gradient/subgradient of it own utility function. An alternative performance analysis is proposed by [9] in a linear regression perspective.

In this paper, we consider a network composed by $N \geq 1$ agents. A given continuously differentiable utility function $f_i : \mathbb{R}^d \to \mathbb{R}$ is associated to each agent $i = 1, \ldots, N$, where d is an integer. We investigate the following minimization problem:

$$\min_{\theta\in G} \sum_{i=1}^{N} f_i(\theta) \tag{1}$$

where G is a subset of $\mathbb{R}^d$ supposed to be known by each agent. We are interested in two distinct cases: first the case of unconstrained minimization ($G = \mathbb{R}^d$), second, the case where G is a compact convex subset specified by inequality constraints. However, we do **not** suppose that the objective function $f := \sum_i f_i$ is convex. Moreover, we consider the context of stochastic approximation: each agent observes a random sequence of noisy observations of the gradient ∇f_i. We are interested in *on-line* estimates of local solutions to (1) using a distributed R-M algorithm.

Our contribution is the following. A distributed R-M algorithm is introduced following [5, 7, 8]. It is proved to converge to a consensus with probability one (w.p.1.) that is, all agents eventually reach an agreement on their estimate of the local solution to the minimization problem (1). In addition, each agent's estimate converges to the set of

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

VALUETOOLS 2011, May 16-20, Paris, France

DOI 10.4108/icst.valuetools.2011.245895

Kuhn-Tucker points $\mathcal{L}_{KT}$ of (1) under some assumptions. In the unconstrained case, the proof is based on the existence of a well-behaved Lyapunov function which ensures the stability of the algorithm. In the constrained case, the proof relies on recent results of [10] about perturbed differential inclusions.

The paper is organized as follows. Section 2 introduces the distributed algorithm and the main assumptions on the network and the observation model. In Section 3, we analyze the behavior of the algorithm in case of unconstrained optimization ($G = \mathbb{R}^d$). We prove the almost sure agreement and the almost sure convergence of the algorithm. We provide the speed of convergence as well as a Central Limit Theorem on the estimates. In Section 4, we investigate the case where the domain G is determined by a set of inequality constraints. Agreement and almost sure convergence to Kuhn-Tucker points is shown. Section 5 provides an example of application to distributed power allocation for ad-hoc wireless networks.

2. THE DISTRIBUTED ALGORITHM

2.1 Description of the Algorithm

Each node i generates a stochastic process $(\theta_{n,i})_{n\geq 1}$ in $\mathbb{R}^d$ using a two-step iterative algorithm:

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

$$\tilde{\theta}_{n,i} = P_G\left[\theta_{n-1,i} + \gamma_n Y_{n,i}\right] \ , \qquad (2)$$

where γ_n is a deterministic positive step size, $Y_{n,i}$ is a random variable, and P_G represents the projection operator onto the set G. In particular, P_G is equal to the identity map in case G is taken to be the whole space $\mathbb{R}^d$. Random variable $Y_{n,i}$ is to be interpreted as a perturbed version of the opposite gradient of f_i at point $\theta_{n-1,i}$. As will be made clear by Assumption **A**1d) below, it is convenient to think of $Y_{n,i}$ as $Y_{n,i} = -\nabla f_i(\theta_{n-1,i}) + \delta M_{n,i}$ where $(\delta M_{n,i})_n$ is a martingale increment sequence which stands for the random perturbation.

`[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.

We cast this algorithm into a more compact vector form. Define the random vectors $\boldsymbol{\theta}_n$ and Y_n as $\boldsymbol{\theta}_n := (\theta_{n,1}^T, \ldots, \theta_{n,N}^T)^T$ and $Y_n = (Y_{n,1}, \ldots, Y_{n,N})^T$ where ^{T} denotes transposition. The algorithm reduces to:

$$\boldsymbol{\theta}_n = (W_n \otimes I_d) P_{G^N}\left[\boldsymbol{\theta}_{n-1} + \gamma_n Y_n\right] \qquad (3)$$

where $\otimes$ denotes the Kronecker product, I_d is the $d \times d$ identity matrix and P_{G^N} is the projector onto the Nth order product set $G^N := G \times \cdots \times G$.

2.2 Observation and Network Models

The time-varying communication network between the nodes is represented by the sequence of random matrices $(W_n)_{n\geq 1}$. For any $n \geq 1$, we introduce the σ-field $\mathcal{F}_n = \sigma(\boldsymbol{\theta}_0, Y_{1:n}, W_{1:n})$. The distribution of the random vector Y_{n+1} conditionally to $\mathcal{F}_n$ is assumed to be such that:

$$\mathbb{P}\left(Y_{n+1} \in A \,|\, \mathcal{F}_n\right) = \mu_{\boldsymbol{\theta}_n}(A)$$

for any measurable set A, where $(\mu_{\boldsymbol{\theta}})_{\boldsymbol{\theta}\in\mathbb{R}^{dN}}$ is a given family of probability measures on $\mathbb{R}^{dN}$. For any $\boldsymbol{\theta} \in \mathbb{R}^{dN}$, define $\mathbb{E}_{\boldsymbol{\theta}}[g(Y)] := \int g(y)\mu_{\boldsymbol{\theta}}(dy)$. Denote by $\mathbb{1}$ the $N \times 1$ vector whose components are all equal to one. Denote by $|x|$ the Euclidean norm of any vector x. It is assumed that:

Assumption 1. *The following conditions hold:*
a) Matrix W_n is doubly stochastic: $W_n\mathbb{1} = W_n^T\mathbb{1} = \mathbb{1}$.
b) $(W_n)_{n\geq 1}$ is a sequence of square-integrable matrix-valued random variables. The spectral radius ρ_n of matrix $\mathbb{E}(W_nW_n^T) - \mathbb{1}\mathbb{1}^T/N$ satisfies:

$$\lim_{n\to\infty} n(1-\rho_n) = +\infty \ .$$

c) For any positive measurable functions g_1, g_2,

$$\mathbb{E}[g_1(W_{n+1})g_2(Y_{n+1})|\mathcal{F}_n] = \mathbb{E}[g_1(W_{n+1})]\,\mathbb{E}_{\boldsymbol{\theta}_n}[g_2(Y)] \ .$$

d) $\boldsymbol{\theta}_0 \in G^N$ and $\mathbb{E}[|\boldsymbol{\theta}_0|^2] < +\infty$.
e) For any $i = 1, \ldots, N$, f_i is continuously differentiable.
f) For any $\boldsymbol{\theta} = (\theta_1^T, \cdots, \theta_N^T)^T$,

$$\mathbb{E}_{\boldsymbol{\theta}}[Y] = -(\nabla f_1(\theta_i)^T, \cdots, \nabla f_N(\theta_N)^T)^T \ .$$

Condition **A1**a) is satisfied provided that the nodes coordinate their weights. Coordination schemes are discussed in [7, 6]. Due to **A1**b), note that $\rho_n < 1$ as soon as n is large enough. Loosely speaking, Assumption **A1**b) ensures that $\mathbb{E}(W_nW_n^T)$ is close enough to the projector $\mathbb{1}\mathbb{1}^T/N$ on the line $\{t\mathbb{1} : t \in \mathbb{R}\}$. This way, the amount of information exchanged in the network remains sufficient in order to reach a consensus. Condition **A1**c) implies that r.v. W_{n+1} and Y_{n+1} are independent conditionally to the past. In addition, $(W_n)_{n\geq 1}$ forms an independent sequence (not necessarily identically distributed). Condition **A1**f) means that each $Y_{n,i}$ can be interpreted as a 'noisy' version of $-\nabla f_i(\theta_{n-1,i})$. The distribution of the random additive perturbation $Y_{n,i} + \nabla f_i(\theta_{n-1,i})$ is likely to depend on the past through the value of $\boldsymbol{\theta}_{n-1}$, but has a zero mean for any given value of $\boldsymbol{\theta}_{n-1}$.

Assumption 2. *a) The deterministic sequence $(\gamma_n)_{n\geq 1}$ is positive and such that $\sum_n \gamma_n = \infty$.*
b) There exists $\alpha > 1/2$ such that:

$$\lim_{n\to\infty} n^\alpha \gamma_n = 0 \qquad (4)$$

$$\liminf_{n\to\infty} \frac{1-\rho_n}{n^\alpha\gamma_n} > 0 \ . \qquad (5)$$

Note that, when (4) holds true then $\sum_n \gamma_n^2 < \infty$, which is a rather common assumption in the framework of decreasing step size stochastic algorithms [11]. In order to have some insights on (5), consider the case where $1 - \rho_n = a/n^\eta$ and $\gamma_n = \gamma_0/n^\xi$ for some constants $a, \gamma_0 > 0$. Then, a sufficient condition for (5) and **A2**a) is:

$$0 \leq \eta < \xi - 1/2 \leq 1/2 \ .$$

In particular, $\xi \in (1/2, 1]$ and $\eta \in [0, 1/2)$. The case $\eta = 0$ typically correspond to the case where matrices W_n are identically distributed. In this case, $\rho_n = \rho$ is a constant w.r.t. n and our assumptions reduce to: $\rho < 1$. However, matrices W_n are not necessarily supposed to be identically distributed. Our results hold in a more general setting. As a matter of fact, all results of this paper hold true when matrices W_n are allowed to converge to the identity matrix (but at a moderate speed, slower than $1/\sqrt{n}$ in any case). Therefore, matrix W_n may be taken to be the identity matrix with high probability, without any restriction on the results presented in this paper. From a communication network point of view, this means that the exchange of information between agents reduces to zero as $n \to \infty$. This remark has practical consequences in case of wireless networks, where it is often required to limit as much as possible the communication overhead.

3. UNCONSTRAINED OPTIMIZATION

3.1 Framework and Assumptions

In this section, G is taken to be the whole space, so that the algorithm (3) simplifies to:

$$\boldsymbol{\theta}_n = (W_n \otimes I_d)\left(\boldsymbol{\theta}_{n-1} + \gamma_n Y_n\right) \ . \quad (6)$$

Our aim is to study the convergence of the above iterate sequence. Note that sequence $\boldsymbol{\theta}_n$ is not *a priori* supposed to stay in a compact set. Additionally, in most situations, large values of some components of $\boldsymbol{\theta}_{n-1}$ may lead to large values of Y_n. Otherwise stated, one of the main issues in the unconstrained case is to demonstrate the **stability** of the algorithm (6) based on explicit and intelligible assumptions on the objective function f and on the stochastic perturbation.

ASSUMPTION 3. *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 $\theta \in \mathbb{R}^d$, $-\nabla V(\theta)^T \nabla f(\theta) \le 0$.
c) There exists a constant C_1, such that for any $\theta \in \mathbb{R}^d$, $|\nabla V(\theta)|^2 \le C_1(1 + V(\theta))$.
d) For any $M > 0$, the level sets $\{\theta \in \mathbb{R}^d : V(\theta) \le M\}$ are compact.
e) The set $\mathcal{L} := \{\theta \in \mathbb{R}^d : \nabla V(\theta)^T \nabla f(\theta) = 0\}$ is bounded.
f) $V(\mathcal{L})$ has an empty interior.

Assumption **A3**b) means that V is a Lyapunov function for $-\nabla f$. In case of gradient systems obtained from optimization problems such as (1), a Lyapunov function V is usually given by the objective function f itself, or by a composition $\phi \circ f$ of f with a well-chosen increasing map ϕ: Assumption **A3**b) is then trivially satisfied. In this case, the set $\mathcal{L}$ reduces to the roots of ∇f:

$$\mathcal{L} = \{\theta \in \mathbb{R}^d : \nabla f(\theta) = 0\} \ .$$

Assumption **A3** combined with the condition $\sum_n \gamma_n = +\infty$ allows to prove the convergence of the deterministic sequence $t_{n+1} = t_n - \gamma_{n+1} \nabla f(t_n)$ to the set $\mathcal{L}$. When ∇f is unknown and replaced by a stochastic approximation, the limiting behavior of the noisy algorithm is similar under some regularity conditions and under the assumption that the step-size sequence satisfies $\sum_n \gamma_n^2 < \infty$. Assumption **A3**c) implies that

θ	dummy variable in $\mathbb{R}^d$
$\boldsymbol{\theta}$	dummy variable in $\mathbb{R}^{dN}$
$\theta_{n,i}$	estimate at agent i and at time n in $\mathbb{R}^d$
$\boldsymbol{\theta}_n$	vector of the N agents estimates in $\mathbb{R}^{dN}$
$\langle \boldsymbol{\theta}_n \rangle$	average of the agents estimates in $\mathbb{R}^d$
J	projector onto the consensus subspace
$J^\perp \boldsymbol{\theta}_n$	disagreement vector between agents in $\mathbb{R}^{dN}$
f	Aggregate utility function $f = \sum_i f_i$
Y_n	vector of all observations at time n, in $\mathbb{R}^{dN}$
$\mathbb{1}$	Vector $(1, \cdots, 1)^T$ in $\mathbb{R}^N$
γ_n	step size
ρ_n	spectral radius of $\mathbb{E}(W_n W_n^T) - \mathbb{1}\mathbb{1}^T/N$

Table 1: Summary of useful notations

V increases at most at quadratic rate $O(|\theta|^2)$ when $|\theta| \to \infty$. Assumption **A3**f) is trivially satisfied when $\mathcal{L}$ is finite.

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

$$\langle \boldsymbol{\theta} \rangle := \frac{1}{N}(\mathbb{1}^T \otimes I_d)\boldsymbol{\theta} \quad (7)$$

is a vector of $\mathbb{R}^d$ equal to $(\theta_1 + \cdots + \theta_N)/N$ in case we write $\boldsymbol{\theta} = (\theta_1^T, \ldots, \theta_N^T)^T$ for some $\theta_1, \ldots, \theta_N$ in $\mathbb{R}^d$.

ASSUMPTION 4. *There exists a constant C_2, such that for any $\boldsymbol{\theta} = (\theta_1^T, \cdots, \theta_N^T)^T$ in $\mathbb{R}^{dN}$,*

$$\mathbb{E}_{\boldsymbol{\theta}}\left[|Y|^2\right] \le C_2\left(1 + V(\langle \boldsymbol{\theta} \rangle) + |J^\perp \boldsymbol{\theta}|^2\right) \quad (8)$$

$$\left|\nabla f(\langle \boldsymbol{\theta} \rangle) - \frac{1}{N}\sum_{i=1}^N \nabla f_i(\theta_i)\right| \le C_2 |J^\perp \boldsymbol{\theta}| \quad (9)$$

Condition (8) implies that $|\nabla f(\theta)|^2 \le C_2(1 + V(\theta))$. This means that the mean field $\nabla f(\theta)$ cannot increase more rapidly than $O(|\theta|)$ as $|\theta| \to \infty$. Condition (9) is in particular satisfied in case all f_i's are Lipschitz function. Condition (9) ensures that small variations of vector $\boldsymbol{\theta}$ near the consensus space cannot produce large variations of $\sum_i \nabla f_i(\theta_i)$.

3.2 Convergence w.p.1

The disagreement between agents can be quantified through the norm of the vector

$$J^\perp \boldsymbol{\theta}_n = \boldsymbol{\theta}_n - \mathbb{1} \otimes \langle \boldsymbol{\theta}_n \rangle \ .$$

LEMMA 1 (AGREEMENT). *Under* **A1-2**, **A3***a-c) and* **A4**,
i) $J^\perp \boldsymbol{\theta}_n$ converges to zero almost surely (a.s.) as $n \to \infty$.
ii) For any $\beta < 2\alpha$, $\lim_{n\to\infty} n^\beta \mathbb{E}\left[|J^\perp \boldsymbol{\theta}_n|^2\right] = 0$.

Lemma 1 is the key result to characterize the asymptotic behavior of the algorithm. The proof is omitted due to lack of space, but will be presented in an extended version of this paper. Point *i)* means that the disagreement between agents converges almost-surely to zero. Point *ii)* states that the convergence also holds in L^2 and that the convergence speed is faster than $1/\sqrt{n}$: This point will be revealed especially useful in Section 3.3. Define $\mathsf{d}(\theta, A) := \inf\{|\theta - \varphi| : \varphi \in A\}$ for any $\theta \in \mathbb{R}^d$ and $A \subset \mathbb{R}^d$. Define $\mathbb{1} \otimes \mathcal{L} := \{\mathbb{1} \otimes \theta : \theta \in \mathcal{L}\}$.

THEOREM 1. *Assume* **A1**, **A2**, **A3** *and* **A4**. *Then, w.p.1,*

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

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 $\langle \boldsymbol{\theta}_n \rangle$ of the network converges to some connected component of $\mathcal{L}$. When $\mathcal{L}$ is finite, Theorem 1 implies that $\boldsymbol{\theta}_n$ converges a.s. to some point in $\mathbb{1} \otimes \mathcal{L}$.

The proof of Theorem 1 is omitted. Conditions **A2**, **A3**a-e) and **A4** 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 $\mathcal{L}$.

3.3 Central Limit Theorem

Let θ_* be a point satisfying the following Assumption.

ASSUMPTION 5. *a)* $\theta_* \in \mathcal{L}$.
b) Function f *is two times differentiable at point* θ_* *and* $f(\theta) = H(\theta_*)(\theta - \theta_*) + O(|\theta - \theta_*|^2)$ *for any* θ *in a neighborhood of* θ_*, *where* $H(\theta_*)$ *denotes the* $d \times d$ *Hessian matrix of* f *at point* θ_*.
c) $H(\theta_*)$ *is a stable matrix: the largest real part of its eigenvalues is* $-L$, *where* $L > 0$.
d) There exists $\delta > 0$ *such that the function:* $\boldsymbol{\theta} \mapsto \mathbb{E}_{\boldsymbol{\theta}}\left[|Y|^{2+\delta}\right]$ *is bounded in a neighborood of* $\mathbb{1} \otimes \theta_*$.
e) The matrix-valued function $Q : \mathbb{R}^{dN} \to \mathbb{R}^{d\times d}$ *defined by:*

$$Q(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\left[(\langle Y \rangle - \mathbb{E}_{\boldsymbol{\theta}} \langle Y \rangle) (\langle Y \rangle - \mathbb{E}_{\boldsymbol{\theta}} \langle Y \rangle)^T \right]$$

is continuous at point $\mathbb{1} \otimes \theta_*$.
f) Matrix $Q(\mathbb{1} \otimes \theta_*)$ *is positive definite.*

ASSUMPTION 6. *For any* $n \geq 1$, $\gamma_n = \gamma_0 \, n^{-\xi}$ *where* $\xi \in (1/2, 1]$ *and* $\gamma_0 > 0$. *In case* $\xi = 1$, *we furthermore assume that* $2L\gamma_0 > 1$.

The normalized disagreement vector $\gamma_n^{-1/2} J^{\perp} \boldsymbol{\theta}_n$ converges to zero in probability by Lemma 1ii). Therefore, it can be shown that the asymptotic analysis reduces to the study of the average $\langle \boldsymbol{\theta}_n \rangle$. To that end, we remark from **A1**a) that $(\mathbb{1} \otimes I_d)(W_n \otimes I_d) = (\mathbb{1} \otimes I_d)$. Thus, $\langle \boldsymbol{\theta}_n \rangle$ satisfies: $\langle \boldsymbol{\theta}_n \rangle = \langle \boldsymbol{\theta}_{n-1} \rangle + \gamma_n \langle Y_n \rangle$. The main step is to rewrite the above equality under the form:

$$\langle \boldsymbol{\theta}_n \rangle = \langle \boldsymbol{\theta}_{n-1} \rangle + \gamma_n \left(-\nabla f(\langle \boldsymbol{\theta}_{n-1} \rangle) + \delta \tilde{M}_n + r_n \right) \ ,$$

where $\delta \tilde{M}_n$ is a martingale increment sequence satisfying some desired properties (details are omitted) and where r_n is a random sequence which is proved to be negligible. The final result is a consequence of [12]. A sequence of r.v. $(X_n)_n$ is said to converge in distribution (stably) to a r.v. X given an event E whenever $\mathbb{E}\left(g(X_n)\mathbb{1}_E\right) = \mathbb{E}\left(g(X)\right)\mathbb{P}(E)$ for any bounded continuous function g.

THEOREM 2. *Assume* **A1**–**4**, **A6** *and assume that there exists a point* θ_* *satisfying* **A5**. *Then, given the event*

$$\{\lim_{n\to\infty} \langle \boldsymbol{\theta}_n \rangle = \theta_*\} \ ,$$

the following holds true:

$$\gamma_n^{-1/2} \left(\boldsymbol{\theta}_n - \mathbb{1} \otimes \theta_* \right) \xrightarrow{\mathcal{D}} \mathbb{1} \otimes Z \ .$$

where Z *is a* $d \times 1$ *zero mean Gaussian vector whose covariance matrix* Σ *is the unique solution to:*

$$\left(H(\theta_*) + \zeta I_d\right) \Sigma + \Sigma \left(H(\theta_*) + \zeta I_d\right) = -Q(\mathbb{1} \otimes \theta_*) \qquad (10)$$

where $\zeta = 0$ *if* $\xi \in (1/2, 1)$ *and* $\zeta = 1/(2\gamma_0)$ *if* $\xi = 1$.

Theorem 2 states that, given the event that sequence $\boldsymbol{\theta}_n$ converges to a given point $\mathbb{1} \otimes \theta_*$, the normalized error $\gamma_n^{-1/2}(\boldsymbol{\theta}_n - \mathbb{1} \otimes \theta_*)$ converges to a Gaussian vector. The latter limiting random vector belongs to the consensus subspace *i.e.*, it has the form $\mathbb{1} \otimes Z$, where Z is a Gaussian r.v. of dimension d. Theorem 2 has the following important consequences. First, thanks to the gossip step, the component of the error vector in the orthogonal consensus subspace is asymptotically negligible. The dominant source of error is due to the presence of observation noise in the algorithm, and not on possible disagreements between agents. As a matter of fact, the limiting behavior of the average estimate is similar to the one that would have been observed in a centralized setting. Interestingly, this remark is true even if the agents reduce their cooperation as time increases (consider the case where $W_n = I_d$ with probability converging to one).

3.4 Influence of the network topology

To illustrate our claims, assume for simplicity that $(W_n)_{n\geq 1}$ is an i.i.d. sequence. Then $\rho_n =: \rho$ is a constant w.r.t. n. In this case, all our hypotheses on sequence $(W_n)_{n\geq 1}$ reduce to:

$$\rho < 1 \ . \qquad (11)$$

In order to have more insights, it is useful to relate the above inequality to a connectivity condition on the network. To that end, we focus on an example. Assume for instance that matrices W_n follow the now widespread asynchronous random pairwise gossip model described in [6]. At a given time instant n, a node i, picked at random, wakes up and exchange information with an other node j also chosen at random (other nodes $k \notin \{i, j\}$ do not participate to any exchange of information). W_n belongs to the alphabet $\{\mathsf{W}_{i,j} : i, j = 1, \ldots, N\}$ where:

$$\mathsf{W}_{i,j} := I_d - (e_i - e_j)(e_i - e_j)^T/2 \ ,$$

where e_i represents the ith vector of the canonical basis ($e_i(k) = 1$ if $i = k$, zero otherwise). Denote by $P_{i,j} = P_{j,i}$ the probability that the active pair of nodes at instant n coincides with the pair $\{i, j\}$. In practice, $P_{i,j}$ is nonzero only if nodes i, j are able to communicate (*i.e.* they are connected). Consider the weighted nondirected graph $\mathcal{G} =$

$(\mathcal{E}, \mathcal{V}, \mathcal{W})$ where $\mathcal{E}$ is the set of vertices $\{1, \dots, N\}$, $\mathcal{V}$ is the set of edges (by definition, i is connected to j iff $P_{i,j} > 0$), and $\mathcal{W}$ associates the weight $P_{i,j}$ to the connected pair $\{i, j\}$. Using [6], it is straightforward to show that condition (11) is equivalent to the condition that $\mathcal{G}$ is connected.

COROLLARY 1. *Replace conditions (1) and (5) with the assumption that $\mathcal{G}$ is connected. Then Theorems 1 and 2 still hold true.*

In particular, the (nonzero) spectral gap of the Laplacian of $\mathcal{G}$ has no impact on the asymptotic behavior of sequence $\boldsymbol{\theta}_n$. Stated differently, the dominant source of error in the asymptotic regime is due to the observation noise. The disagreement between agents is negligible even in networks with a low level of connectivity.

4. CONSTRAINED OPTIMIZATION

4.1 Framework and Assumptions

We now study the case where the set G is determined by a set of p inequality constraints ($p \geq 1$):

$$G := \left\{\theta \in \mathbb{R}^d \ : \ \forall j = 1, \dots, p, \ q_j(\theta) \leq 0\right\} \tag{12}$$

for some functions $q_1, \dots, q_p$ which satisfy the following conditions. Denote by ∂G the boundary of G. For any $\theta \in G$, we denote by $A(\theta) \subset \{1, \dots, p\}$ the set of active constraints *i.e.*, $q_j(\theta) = 0$ if $j \in A(\theta)$ and $q_j(\theta) < 0$ otherwise.

ASSUMPTION 7. *a) The set G defined by (12) is compact.*
b) For any $j = 1, \dots, p$, $q_j : \mathbb{R}^d \to \mathbb{R}$ is a convex function
c) For any $j = 1, \dots, p$, q_j is two times continuously differentiable in a neigborhood of ∂G.
c) For any $\theta \in \partial G$, $\{\nabla q_j(\theta) : j \in A(\theta)\}$ is a linearly independent collection of vectors.

In the particular case where all utility functions $f_1, \dots, f_N$ are assumed convex, it is possible to study the convergence w.p.1 of the algorithm (3) following an approach similar to [7], and to prove under some conditions that consensus is achieved at a global minimum of the aggregate objective function f. Nevertheless, utility functions may not be convex in a large number of situations, and there seems to be few hope to generalize the proof of [7] in such a wider setting. In this paper, we do **not** assume that the utility functions are convex. In this situation, convergence to a global minimum of (1) is no longer guaranteed. We nevertheless prove the convergence of the algorithm (3) to the set of Kuhn-Tucker (KT) points $\mathcal{L}_{KT}$:

$$\mathcal{L}_{KT} := \{\theta \in G \ : \ -\nabla f(\theta) \in \mathcal{N}_G(\theta)\} \ ,$$

where $\mathcal{N}_G(\theta)$ is the normal cone to G *i.e.*, $\mathcal{N}_G(\theta) := \{v \in \mathbb{R}^d \ : \ \forall \theta' \in G, v^T(\theta - \theta') \geq 0\}$. To prove convergence, we need one more hypothesis:

ASSUMPTION 8. *The following two conditions hold:*
a) $\sup_{\boldsymbol{\theta} \in G^N} \mathbb{E}_{\boldsymbol{\theta}}[\,|Y|^2] < \infty$.
b) Inequality (9) holds for any $\boldsymbol{\theta} \in G^N$.

4.2 Convergence w.p.1

Theorem 3 below establishes two points: First, a consensus is achieved as n tends to infinity, meaning that $J^\perp \boldsymbol{\theta}_n$ converges a.s. to zero. Second, the average estimate $\langle \boldsymbol{\theta}_n \rangle$ converges to the set of KT points.

THEOREM 3. *Assume* **A1**, **A2**, **A7** *and* **A8**. *Then, w.p.1,*

$$\lim_{n\to\infty} \mathsf{d}(\boldsymbol{\theta}_n, \mathbb{1} \otimes \mathcal{L}_{KT}) = 0 \ .$$

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

As a consequence, if $\mathcal{L}_{KT}$ contains only isolated points, sequence $\langle \boldsymbol{\theta}_n \rangle$ converges almost surely to one of these points. The complete proof of Theorem 3 is omitted. We however provide some elements of the proof in the next paragraph.

4.3 Sketch of the proof

To simplify the presentation, we shall focus on the case $p = 1$ *i.e.*, there is only one inequality constraint. We put $q := q_1$ and define $e := \nabla q / |\nabla q|$ the normalized gradient of function q (e is well defined in a neighborhood of ∂G by **A7**c)).

Step 1: *Agreement is achieved as $n \to \infty$.*
Similarly to the unconstrained optimization case (recall previous Lemma 1), the first step of the proof of Theorem 3 is to establish that $|J^\perp \boldsymbol{\theta}_n|$ converges a.s. to zero. As a noticeable difference with the unconstrained case, here stability issues do not come into play as G is bounded (for this reason, the proof of agreement is simpler than in the unconstrained case).

Step 2: *Expression of the average $\langle \boldsymbol{\theta}_n \rangle$ in a R-M like form.*
Using $(\mathbb{1} \otimes I_d)(W_n \otimes I_d) = (\mathbb{1} \otimes I_d)$, it is convenient to write $\langle \boldsymbol{\theta}_n \rangle = \langle \boldsymbol{\theta}_{n-1} \rangle + \gamma_n Z_n$ where

$$Z_n := \frac{1}{\gamma_n N} \sum_{i=1}^N P_G(\theta_{n-1,i} + \gamma_n Y_{n,i}) - \theta_{n-1,i} \ .$$

Consider the martingale increment sequence $\Delta_n := Z_n - \mathbb{E}(Z_n | \mathcal{F}_{n-1})$. From Assumption **A8**a), it can be shown that $\sup_n \mathbb{E}[\,|\Delta_n|^2] < \infty$. Now note that for any $\theta \in G$, $y \in \mathbb{R}^d$,

$$\lim_{\gamma \downarrow 0} \gamma^{-1} \left(P_G(\theta + \gamma y) - \theta\right) = y - (y^T e(\theta))^+ e(\theta) \mathbf{1}_{\partial G}(\theta) \ , \tag{13}$$

where $(x)^+ := \max(x, 0)$ and where $\mathbf{1}_{\partial G}$ is the indicator function of ∂G. Using (13) along with **A7**c) and **A8**b) and the fact that $|J^\perp \boldsymbol{\theta}_n|$ converges to zero, we obtain after some algebra:

$$\langle \boldsymbol{\theta}_n \rangle = \langle \boldsymbol{\theta}_{n-1} \rangle + \gamma_n h(\boldsymbol{\theta}_{n-1}) + \gamma_n \Delta_n + \gamma_n u_n \tag{14}$$

where u_n is some sequence which converges to zero a.s. and where we defined for any $\boldsymbol{\theta} \in G^N$,

$$h(\boldsymbol{\theta}) := -\nabla f(\langle \boldsymbol{\theta} \rangle) - \frac{e(\langle \boldsymbol{\theta} \rangle)}{N} \sum_{i=1}^N \mathbb{E}_{\boldsymbol{\theta}} \left[(Y_i^T e(\theta_i))^+\right] \mathbf{1}_{\partial G}(\theta_i) \ .$$

Step 3: *From equality to inclusion.*
Equality (14) is still far from a conventional R-M equation. Indeed, the second term of the righthand side $\gamma_n h(\boldsymbol{\theta}_{n-1})$ is not a function of the average $\langle \boldsymbol{\theta}_{n-1} \rangle$ as it depends on the

whole vector $\boldsymbol{\theta}_n$. Of course, since the agreement is achieved for large n, $\boldsymbol{\theta}_{n-1}$ should be close to $\mathbb{1} \otimes \langle \boldsymbol{\theta}_{n-1} \rangle$. If h were continuous, one could thus write $h(\boldsymbol{\theta}_{n-1}) \simeq h(\mathbb{1} \otimes \langle \boldsymbol{\theta}_{n-1} \rangle)$ solving this way the latter issue. This is unfortunately not the case, due to the presence of indicator functions in the definition of h. We must resort to inclusions. For any $\epsilon \geq 0$ and any $\theta \in G$, define the following subset of $\mathbb{R}^d$:

$$F_\epsilon(\theta) := \left\{ -\nabla f(\theta) - x\, e(\theta) \mathbf{1}_{\mathsf{d}(\theta, \partial G) \leq \epsilon} \ :\ x \in [0, M] \right\}$$

where $M < \infty$ is a fixed constant chosen as large as needed, and where $\mathbf{1}_{\mathsf{d}(\theta,\partial G)\leq\epsilon}$ is equal to one if θ is at distance less than ϵ of the boundary, and to zero otherwise. In particular, $\mathbf{1}_{\mathsf{d}(\theta,\partial G)\leq\epsilon} = \mathbf{1}_{\partial G}(\theta)$ for $\epsilon = 0$. It is straightforward to show that:

$$\forall \boldsymbol{\theta} \in G^N,\ h(\boldsymbol{\theta}) \in F_{|J^\perp \boldsymbol{\theta}|}(\langle \boldsymbol{\theta} \rangle)$$

provided that M is chosen large enough. Finally, equality (14) can be interpreted in terms of the following inclusion:

$$\langle \boldsymbol{\theta}_n \rangle \in \langle \boldsymbol{\theta}_{n-1} \rangle + \gamma_n F_{\epsilon_n}(\langle \boldsymbol{\theta}_{n-1} \rangle) + \gamma_n \Delta_n + \gamma_n u_n \qquad (15)$$

where we defined for simplicity $\epsilon_n := |J^\perp \boldsymbol{\theta}_{n-1}|$.

Step 4: *Interpolated process and differential inclusions.* From this point to the end of the proof, we shall now study one fixed trajectory $(\langle \boldsymbol{\theta}_n(\omega) \rangle)_n$ of the random process $\langle \boldsymbol{\theta}_n \rangle$, where ω belongs to an event of probability one such that $\epsilon_n(\omega) \to 0$, $u_n(\omega) \to 0$ as n tends to infinity, and sequence $(\Delta_n(\omega))_n$ satisfies some asymptotic rate of change condition (see [11, 10] for details). Dependencies in ω are however omitted for simplicity. Motivated by the approach of [10], we consider the following continuous-time interpolated process. Define $\tau_n = \sum_{k=1}^n \gamma_k$ and

$$\Theta(t) := \langle \boldsymbol{\theta}_{n-1} \rangle + \frac{\langle \boldsymbol{\theta}_n \rangle - \langle \boldsymbol{\theta}_{n-1} \rangle}{\tau_n - \tau_{n-1}} (t - \tau_n)\,, \quad \tau_{n-1} \leq t < \tau_n\ .$$

The next step is to prove that Θ is a *perturbed solution* to the differential inclusion:

$$\frac{dx(t)}{dt} \in F_0(x(t))\ . \qquad (16)$$

When we write that x is a solution to (16), we mean that x is an absolutely continuous mapping $x : \mathbb{R} \to \mathbb{R}^d$ such that (16) is satisfied for almost all $t \in \mathbb{R}$. A function Θ is a perturbed solution to (16) if it 'shadows' the behavior of a solution to (16) as $t \to \infty$ in a sense made clear in [10]. In order to prove that Θ is a perturbed solution to (16), the materials are close to those of [10] (see Proposition 1.3) with some care, however, about the fact that the mean field F_{ϵ_n} is nonhomogeneous in our context (it depends on time n). The proof is concluded by straightforward application of [10]. Consider the differential inclusion (16): function f is a Lyapunov function for the set of KT points $\mathcal{L}_{KT}$. Therefore, by [10], the limit set

$$\bigcap_{t\geq 0} \overline{\Theta([t, +\infty))}$$

is included in $\mathcal{L}_{KT}$. This concludes the proof.

5. APPLICATION: POWER ALLOCATION

5.1 Framework

The context of power allocation for wireless networks has recently raised a great deal of attention in the field of distributed optimization, cooperative and noncooperative game theory (see [13] and references therein). We consider an *ad hoc* network composed of N transmit-destination pairs. Each agent/user sends digital data to its receiver through K parallel (sub)channels. The channel gain of the ith user at the kth subchannel is represented by a positive coefficient $A^{i,i;k}$ which can be interpreted as the square modulus of the corresponding complex valued channel gain. As all agents share the same spectral band, user i suffers from the multiuser interference produced by other users $j \neq i$. Denote by $p^{i;k} \geq 0$ the power allocated by user i to the qth subchannel. We assume that $\sum_{k=1}^K p^{i;k} \leq \mathcal{P}_i$ where $\mathcal{P}_i$ is the maximum allowed power for user i. Define $p^i = [p^{i;1}, \cdots, p^{i;K}]^T$ and $\theta = [p^{1^T}, \cdots, p^{N^T}]^T$ the vector of all powers of all users of size $d := KN$. Assuming deterministic channels, user i is able to provide its destination with rate $R_i(\theta)$ given by (see *e.g.* [14])

$$R_i(\theta, A^i) = \sum_{k=1}^K \log\left(1 + \frac{A^{i,i;k} p^{i;k}}{\sigma_i^2 + \sum_{j\neq i} A^{j,i:k} p^{j;k}}\right)$$

where $A^{j,i:k}$ is the (positive) channel gain between transmitter j and the destination of the ith transmit-destination pair, and where $A^i = [A^{1,i;1}, \cdots, A^{N,i:K}]^T$. Here, σ_i^2 is the variance of the additive white Gaussian noise at the destination of source i. The aim is to select a relevant value for the resource allocation parameter $\theta \in G$ in a distributed fashion, where G is the set of constraints obtained from the aforementioned power constraints $\mathcal{P}_1, \ldots, \mathcal{P}_N$ and positivity constraints.

5.2 Deterministic Coalitional Allocation

To simplify the presentation, we first consider the case of fixed deterministic channel gains $A^1, \ldots, A^N$. A widespread approach consists in computing θ through the so-called best response dynamics. At every step of the iteration, an agent i updates its own power vector p^i assuming other users' power to be fixed. This is the well known iterative water filling algorithm [14]. Here, we are interested in a different perspective. The aim is rather to search for social fairness between users. We aim at finding a local maximum of the following weighted sum rate:

$$\sum_{i=1}^N \beta_i\, R_i(\theta, A^i) \qquad (17)$$

where β_i is an arbitrary positive deterministic weight known only by agent i. Consider the following deterministic gradient algorithm. Each user i has an estimate $\theta_{n,i}$ of θ at the nth iteration. Here, we stress the fact that a given user has not only an estimate of what should be its own power allocation p^i, but has also an estimate of what should be the power allocation of other users $j \neq i$. Denote by $\boldsymbol{\theta}_n = [\theta_{n,1}^T, \cdots, \theta_{n,N}^T]^T$ the vector of size $dN = KN^2$ which gathers all local estimates. Similarly to (3), a distributed algorithm for the maximization of (17) would have the form $\boldsymbol{\theta}_n = (W_n \otimes I_d) P_{G^N}\,[\boldsymbol{\theta}_{n-1} + \gamma\, Y(\boldsymbol{\theta}_{n-1}; A)]$ where $Y(\boldsymbol{\theta}; A) = [\beta_1 \nabla_\theta R_i(\theta_1; A^1)^T, \cdots, \beta_N \nabla_\theta R_i(\theta_N; A^N)^T]^T$ and where ∇_θ

is the gradient operator with respect to the first argument θ of $R_i(\theta, A^i)$.

5.3 Stochastic Coalitional Allocation

In many situations however, the above algorithm is impractical. This is for instance the case when the channel gains are random and rapidly time-varying in an ergodic fashion. This is also the case when channel gains are known only up to a random perturbation. In such settings, it more likely that each user i observes a random sequence $(A^i_n)_{n\geq 1}$, where $A^1_n, \dots, A^N_n$ typically correspond to the realization at time n of a time-varying ergodic channel. The distributed optimization scheme is given by equation (3) where

$$Y_{n,i} = \beta_i \nabla_\theta R_i(\theta_{n,i}; A^i_n) \ .$$

Assume for the sake of simplicity that sequence $(A^1_n, \dots, A^N_n)_n$ is i.i.d. By Theorem 3, all users converge to a consensus on the global resource allocation parameters. After convergence of the distributed R-M algorithm, the resource allocation parameters achieve a Kuhn-Tucker point of the optimization problem:

$$\max_{\theta \in G} \sum_{i=1}^{N} \beta_i \mathbb{E}[R_i(\theta, A^i)] \tag{18}$$

where the expectation in the inner sum is taken w.r.t. the channel coefficients A^i.

We provide some numerical results. Consider four nodes: 1 is connected to 2 ($1 \sim 2$), $1 \sim 3$, $2 \sim 3$, $2 \sim 4$, $3 \sim 4$. Assume $Q = 2$, $\beta_1 = \beta_3 = 0.3$, $\beta_2 = \beta_4 = 0.2$, $\sigma_1^2 = \sigma_4^2 = 0.1$, $\sigma_2^2 = 0.05$, $\sigma_3^2 = 0.02$. Assume that all r.v. $A_n^{i,j;k}$ are i.i.d. with standard exponential distribution. The algorithm is initialized at a random point $\boldsymbol{\theta}_0$. Figure 1 illustrates the fact that the disagreement between agents $|J^\perp \boldsymbol{\theta}_n|$ converges to zero as n tends to infinity. Figure 2 shows the estimated value of the objective function given by (18). The expectation in (18) is estimated using 10^3 Monte-Carlo trials at each iteration.

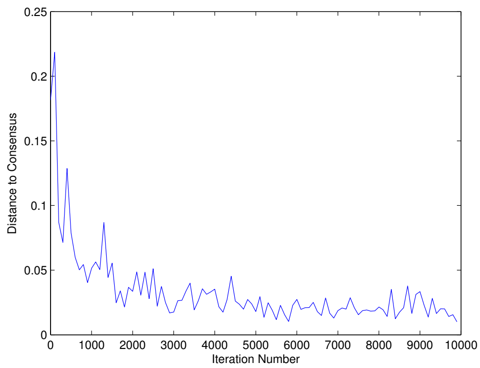

Figure 1: $|J^\perp \boldsymbol{\theta}_n|$ as a function of n.

Acknowledgements

The Authors would like to thank Gersende Fort and Walid Hachem for their contribution to this work. They are as well grateful to Eric Moulines for fruitful discussions.

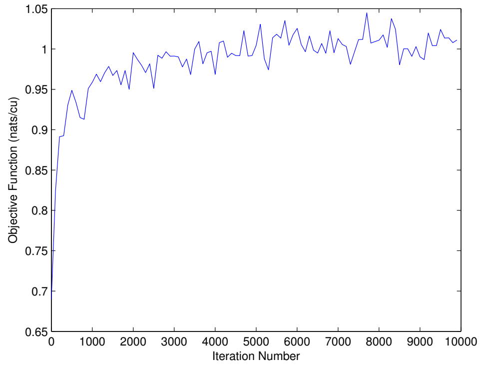

Figure 2: Estimated value of the objective function.

6. REFERENCES

[1] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. of Mathem. Statist.*, vol. 22, no. 3, pp. 400–407, 1951.

[2] B. Delyon, M. Lavielle, and E. Moulines, "Convergence of a stochastic approximation version of the EM algorithm," *Annals of Statistics*, vol. 27, no. 1, pp. 94–128, 1999.

[3] S. Gadat and L. Younes, "A stochastic algorithm for feature selection in pattern recognition," *The Journal of Machine Learning Research*, vol. 8, pp. 509–547, 2007.

[4] B. Widrow, J.M. McCool, M.G. Larimore, and C.R. Johnson Jr, "Stationary and nonstationary learning characteristics of the LMS adaptive filter," *Proceedings of the IEEE*, vol. 64, no. 8, pp. 1151–1162, 1976.

[5] Bertsekas D. P. and Tsitsiklis J. N., *Parallel and distributed computation: numerical methods*, Athena Scientific, first edition edition, 1997.

[6] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized gossip algorithms," *IEEE Trans. on Inform. Theory*, vol. 52, no. 6, pp. 2508–2530, 2006.

[7] A. Nedic, A. Ozdaglar, and P.A. Parrilo, "Constrained consensus and optimization in multi-agent networks," *IEEE Trans. on Automatic Control*, vol. 55, no. 4, pp. 922–938, April 2010.

[8] P. Bianchi, G. Fort, W. Hachem, and J. Jakubowicz, "On the convergence of a distributed parameter estimator for sensor networks with local averaging of the estimate," in *ICASSP*, Praha, Czech Republic, 2011.

[9] 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*, 2008, pp. 1535–1540.

[10] M. Benaim, J. Hofbauer, and S. Sorin, "Stochastic approximations and differential inclusions," *SIAM Journal on Control and Optimization*, vol. 44, no. 1, pp. 328–348, 2005.

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

[12] M. Pelletier, "Weak convergence rates for stochastic approximation with application to multiple targets and simulated annealing," *Annals of Applied Probability*, vol. 8, no. 1, pp. 10–44, 1998.

[13] P. Mertikopoulos, E.V. Belmega, A. Moustakas, and S. Lasaulce, "Distributed learning policies for power allocation in multiple access channels," *submitted to IEEE Journal of Selected Areas in Communications*, 2011.

[14] G. Scutari, D.P. Palomar, and S. Barbarossa, "The MIMO iterative waterfilling algorithm," *IEEE Transactions on Signal Processing*, vol. 57, no. 5, pp. 1917–1935, May 2009.