PERFORMANCE ANALYSIS OF A DISTRIBUTED ROBBINS-MONRO ALGORITHM FOR SENSOR NETWORKS

Pascal Bianchi, Gersende Fort, Walid Hachem, Jérémie Jakubowicz

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

ABSTRACT

This paper investigates the rate of convergence of a distributed Robbins-Monro algorithm for sensor networks. The algorithm under study consists of two steps: a local Robbins-Monro step at each sensor and a gossip step that drives the network to a consensus. Under verifiable sufficient conditions, we give an explicit rate of convergence for this algorithm and provide a conditional Central Limit Theorem. Our results are applied to distributed source localization.

1. INTRODUCTION

The Robbins-Monro (R-M) algorithm [1] is a widely used procedure for finding the roots of an unknown function h. Its applications range from Statistics (e.g. [2]) to Electrical Engineering (e.g. [3]) through Communication Networks (e.g. [4]) and Machine Learning (e.g. [5]). Formally the R-M algorithm can be summarized as an iterative scheme of the form $\theta_{n+1} = \theta_n + \gamma_{n+1}(h(\theta_n) + \xi_{n+1})$ where the sequence $(\theta_n)_{n\in\mathbb{N}}$ will eventually converge to a zero of h and 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 (e.g., [6, 7]). 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 distributed fashion is to use gossip algorithms [8, Chapter 7]. Recently, [9, 10] considered a distributed R-M algorithm combining the algorithm of [8] with a random gossip approach[11]. In [12], we address the stability and the convergence of this stochastic algorithm. In this paper, we provide an analysis of the rate of convergence: we prove a conditional Central Limit Theorem (CLT) under explicit assumptions. To the best of our knowledge, this CLT type result is novel for such distributed R-M algorithms.

The paper is organized as follows. Section 2 is devoted to a detailed description of the proposed algorithm. In Section 3 we state a set of sufficient assumptions for convergence. The almost sure convergence of the algorithm is studied in Section 4. Section 5 contains the main contribution of this paper, a central limit theorem. In Section 6, the algorithm is applied to distributed source localization.

2. DISTRIBUTED ALGORITHM

Consider a network composed by $N \geq 1$ nodes, and assume that node $i \in \{1, \ldots, N\}$ 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-step 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 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 $n \geq 1$, $i \in \{1, \ldots, N\}$, $X_{n,i} \in \mathsf{X}$ where X represents an arbitrary measurable space. Define the function $H : \mathbb{R}^{dN} \times \mathsf{X}^N \to \mathbb{R}^{dN}$ 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, $x = (x_1, \ldots, x_N)^T$ 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 $X_n = (X_{n,1}, \ldots, X_{n,N})^T$. 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 Kronecker product and I_d is the $d \times d$ identity matrix.

3. ASSUMPTIONS

3.1 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 $\mathcal{F}_n = \sigma(\boldsymbol{\theta}_0, \bar{X}_{1:n}, W_{1:n})$. The distribution of random vector X_{n+1} conditionally to $\mathcal{F}_n$ is assumed to be such that:

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

for any measurable set $A \subset \mathsf{X}^N$, where $(\mu_{\boldsymbol{\theta}})_{\boldsymbol{\theta}\in\mathbb{R}^{dN}}$ is a given family of probability measures on X^N. For any $\boldsymbol{\theta} \in \mathbb{R}^{dN}$, set $\mathbb{E}_{\boldsymbol{\theta}}[g(X)] := \int g(x)\mu_{\boldsymbol{\theta}}(dx)$. 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 A1.
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 f, g,

$$\mathbb{E}[f(W_{n+1})g(X_{n+1})|\mathcal{F}_n] = \mathbb{E}[f(W_{n+1})]\,\mathbb{E}_{\boldsymbol{\theta}_n}[g(X)] \ .$$

d) $\mathbb{E}[|\boldsymbol{\theta}_0|^2] < +\infty$.

Condition **A1**a) is satisfied provided that the nodes coordinate their weights. Coordination schemes are discussed in [9, 11]. The condition also holds in case of asynchronous networks (see [11] for details). 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 X_{n+1} are independent conditionally to the past. In addition, $(W_n)_{n\geq 1}$ form an independent sequence.

3.2 Further 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

$$\langle\boldsymbol{\theta}\rangle := \frac{1}{N}(\mathbb{1}^T \otimes I_d)\boldsymbol{\theta} \tag{3}$$

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 $\mathbb{R}^d$. We introduce the *mean field* of the distributed Robbins-Monro algorithm as the function $h : \mathbb{R}^d \to \mathbb{R}^d$ given by:

$$h(\theta) := \mathbb{E}_{\mathbb{1}\otimes\theta}\left[\langle H(\mathbb{1} \otimes \theta; X)\rangle\right] \ , \tag{4}$$

where $\langle H(\boldsymbol{\theta}; x)\rangle = \frac{1}{N}(\mathbb{1}^T \otimes I_d)H(\boldsymbol{\theta}; x)$ is the average of $H(\boldsymbol{\theta}; x)$ (see Eq.(3)). Finally, ∇ is the gradient operator.

3.3 Algorithm Assumptions

We first make some assumptions about the step size of the algorithm.

θ	dummy variable in $\mathbb{R}^d$
$\boldsymbol{\theta}$	dummy variable in $\mathbb{R}^{dN}$
$\theta_{n,i}$	estimate of sensor i at time n in $\mathbb{R}^d$
$\boldsymbol{\theta}_n$	vector of the N sensors' estimates in $\mathbb{R}^{dN}$
$\langle\boldsymbol{\theta}_n\rangle$	average of the sensors estimates in $\mathbb{R}^d$
J	projector onto the consensus subspace
$J^\perp\boldsymbol{\theta}_n$	disagreement vector between sensors in $\mathbb{R}^{dN}$
X_n	vector of all observations at time n, in X^N
$h(\theta)$	mean field of the algorithm $h : \mathbb{R}^d \to \mathbb{R}^d$
$V(\theta)$	Lyapunov function associated with h
$H(\boldsymbol{\theta}; X)$	vector valued function in $\mathbb{R}^{dN} \times \mathsf{X}^N \to \mathbb{R}^{dN}$
$\langle H(\boldsymbol{\theta}; X)\rangle$	average of function H in $\mathbb{R}^{dN} \times \mathsf{X}^N \to \mathbb{R}^d$
$\mathbb{1}$	Vector $(1, \cdots, 1)^T$ in $\mathbb{R}^N$
γ_n	step size
ρ_n	spectral radius of $\mathbb{E}(W_nW_n^T) - \mathbb{1}\mathbb{1}^T/N$

Table 1: Summary of useful notations

Assumption A2.
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:

$$\lim_{n\to\infty} n^\alpha\gamma_n = 0 \tag{5}$$

$$\liminf_{n\to\infty} \frac{1-\rho_n}{n^\alpha\gamma_n} > 0 \ . \tag{6}$$

Note that, when (5) holds true then $\sum_n \gamma_n^2 < \infty$, which is a rather common assumption in the framework of decreasing step size stochastic algorithms [13]. In order to have some insights on (6), 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 (6) 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. Theoretically, matrices W_n are allowed to converge to the identity matrix (but at a moderate speed, slower than $1/\sqrt{n}$ in any case).

Assumption 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 sets $\{\theta \in \mathbb{R}^d : V(\theta) \leq M\}$ are 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 **A3**b) means that V is a Lyapunov function for the mean field h. When h is 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 $\sum_n \gamma_n^2 < \infty$ (sse for instance [13]. Assumption **A3**c) implies that the Lyapunov function 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 assume:

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

$$\mathbb{E}_{\boldsymbol{\theta}} \left[|H(\boldsymbol{\theta}; X)|^2 \right] \leq C_2 \left(1 + V(\langle \boldsymbol{\theta} \rangle) + |J^\perp \boldsymbol{\theta}|^2 \right) \quad (7)$$

$$\mathbb{E}_{\boldsymbol{\theta}} \left| \langle H(\boldsymbol{\theta}; X) \rangle - \langle H(J\boldsymbol{\theta}; X) \rangle \right| \leq C_2 |J^\perp \boldsymbol{\theta}| \quad (8)$$

$$\left| \mathbb{E}_{\boldsymbol{\theta}} \langle H(\boldsymbol{\theta}; X) \rangle - \mathbb{E}_{J\boldsymbol{\theta}} \langle H(J\boldsymbol{\theta}; X) \rangle \right| \leq C_2 |J^\perp \boldsymbol{\theta}| \ . \quad (9)$$

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

Condition (7) implies that $|h(\theta)|^2 \leq C_2(1 + V(\theta))$. This means that the mean field $h(\theta)$ cannot increase more rapidly than $O(|\theta|)$ as $|\theta| \to \infty$. Conditions (8)-(9) are Lipschitz-like conditions which ensure that small variations of vector $\boldsymbol{\theta}$ near the consensus space cannot produce large variations of H.

4. CONSENSUS BETWEEN SENSORS

In this section, we recall some results presented in [12] and extend these results to the case of a non i.i.d. matrix sequence $(W_n)_{n \geq 1}$.

4.1 Agreement between sensors

The disagreement between sensors 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 and recurrence)**.** *Under* **A1**, **A2**, **A3***a-c) and* **A4**,

i) $J^\perp \boldsymbol{\theta}_n$ converges to zero almost surely 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 the lack of space, but will be presented in an extended version of this paper. Point *i)* means that the disagremeent between sensors 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 5.

4.2 Almost sure convergence

Define the distance $\mathsf{d}(\theta, A)$ between a point $\theta \in \mathbb{R}^d$ and a subset $A \subset \mathbb{R}^d$ by $\mathsf{d}(\theta, A) := \inf\{|\theta - \varphi| \ : \ \varphi \in A\}$. Define $\mathbb{1} \otimes \mathcal{L} := \{\mathbb{1} \otimes \theta \ : \ \theta \in \mathbb{R}^d\}$.

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 $\mathcal{L}$.*

The proof 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 $V(\mathcal{L})$.

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 converge to some connected component of $\mathcal{L}$. When $\mathcal{L}$ is finite, Theorem 1 implies that, almost surely, $\boldsymbol{\theta}_n$ converges to some point in $\mathbb{1} \otimes \mathcal{L}$.

5. A CENTRAL LIMIT THEOREM

5.1 Further assumptions

Let θ_* be a point satisfying the following assumption.

Assumption A5. *a) $\theta_* \in \mathcal{L}$.*
b) The mean field h is differentiable at point θ_* and $h(\theta) = \nabla h(\theta_*)(\theta - \theta_*) + O(|\theta - \theta_*|^2)$ for any θ in a neighborhood of θ_*, where $\nabla h(\theta_*)$ denotes the $d \times d$ Jacobian matrix of h at point θ_*.
c) $\nabla h(\theta_*)$ 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:

$$\boldsymbol{\theta} \mapsto \mathbb{E}_{\boldsymbol{\theta}} \left[|H(\boldsymbol{\theta}; X)|^{2+\delta} \right]$$

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:

$$Q(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}} \big[\left(\langle H(\boldsymbol{\theta}, X) \rangle - \mathbb{E}_{\boldsymbol{\theta}} \langle H(\boldsymbol{\theta}, X) \rangle \right) \cdot \left(\langle H(\boldsymbol{\theta}, X) \rangle - \mathbb{E}_{\boldsymbol{\theta}} \langle H(\boldsymbol{\theta}, X) \rangle \right)^T \big]$$

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

Assumption A6.
a) For any $n \geq 1$, $\gamma_n = \gamma_0 \, n^{-\xi}$ where $\xi \in (1/2, 1]$ and $\gamma_0 > 0$.
b) In case $\xi = 1$, we furthermore assume that $2L\gamma_0 > 1$.

5.2 Main result

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

$$\langle \boldsymbol{\theta}_n \rangle = \langle \boldsymbol{\theta}_{n-1} \rangle + \gamma_n \left(h(\langle \boldsymbol{\theta}_{n-1} \rangle) + \epsilon_n + r_n \right) \ ,$$

where ϵ_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 [14, Theorem 1]. A sequence of r.v. $(Y_n)_n$ is said to converge in distribution (stably) to a r.v. Y given an event E whenever $\mathbb{E}\left(f(Y_n)\mathbb{1}_E\right) = \mathbb{E}\left(f(Y)\right)\mathbb{P}(E)$ for any bounded continuous function f.

Theorem 2. *Assume* **A1**–**4**, **A6** *and assume that there exists a point θ_* satisfying* **A5**. *Then, given the event $\{\lim_{n\to\infty}\langle\boldsymbol{\theta}_n\rangle = \theta_*\}$,*

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

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

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

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.

5.3 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:

$$\rho < 1\ . \tag{11}$$

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 pairwaise gossip model described in [11]. 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 $\{1,\ldots,N\}$, $\mathcal{V}$ 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 [11], it is straightforward to show that condition (11) is equivalent to the condition that $\mathcal{G}$ is connected.

Corollary 1. *Replace conditions (1) and (6) 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 sensors is negligible even in networks with a low level of connectivity.

6. NUMERICAL RESULTS

We consider the application of the above algorithm to distributed statistical inference by maximum likelihood. D sources are positioned at unknown locations (in $\mathbb{R}^2$) and these locations are estimated by N sensors. The unknown locations are collected in θ_*, $\theta_*\in\mathbb{R}^d$ where $d = 2D$, and we assume that $X_n\in\mathbb{R}^N$, the received signal energy measurements at each sensor at time n, are i.i.d. r.v. with distribution $\mu(dx_1,\cdots,dx_N) = \prod_{i=1}^N f_{*,i}(x_i)dx_i$ so that $\mu_{\boldsymbol{\theta}} = \mu$ for any $\boldsymbol{\theta}$. We consider a parametric family of density distributions on $\mathbb{R}^N$, $\{f(\cdot;\theta),\theta\in\mathbb{R}^d\}$ of the form $f(x_{1:N};\theta) = \prod_{i=1}^N f_i(x_i;\theta)$ and we want to compute the maximum likelihood estimate. Therefore, in this setting,

$$H_i(\theta;x_i) = \nabla_\theta \log f_i(x_i;\theta)\,,$$

and

$$h(\theta) = \frac{1}{N}\sum_{i=1}^N \int \nabla_\theta \log f_i(x_i;\theta) f_{*,i}(x_i)dx_i\ .$$

It is easily checked that the Lyapunov function V is the Kullback-Leibler divergence between f_* and the density $f(\cdot;\theta)$, so that the above theoretical results show that the sequence of vectors $\{\boldsymbol{\theta}_n, n\geq 0\}$ defined by (2) converges to the set of the stationary points of V. Therefore, this distributed algorithm has the same asymptotic behavior as a centralized Maximum Likelihood algorithm.

In the numerical applications, we consider a network with $N = 40$ sensors and $D = 1$ source. The graph is shown in Figure 1. We run the algorithm described by (2): *(i)* Matrices W_n are chosen as in Section 5.3. *(ii)* the step sequence γ_n is set to $c_1/n^{0.6}$ for $n\leq 10000$ iterations, $c_2(\log n/n)^{0.6}$ for $10000 < n\leq 20000$ and $c_3(\log n/n)^{0.6}$ for $n > 20000$ with $c_1 < c_2 < c_3$. *(iii)* the initial value $\boldsymbol{\theta}_0\in\mathbb{R}^{dN}$ is chosen at random under the uniform distribution on the square $[0,50]\times[0,50]$. The observations are obtained by choosing $f_{*,i}$ as a Gaussian distribution on $\mathbb{R}$ with mean $m_{*,i}$ and variance σ_*^2 given by

$$m_{*,i} = \frac{1000}{|\theta_* - r_i|^2}\,,\quad \sigma_*^2 = \bar{s}^2\,10^{-0.3}$$

where $r_i\in\mathbb{R}^2$ is the location of sensor i and $\bar{s}^2 = N^{-1}\sum_{i=1}^N m_{*,i}^2$. Finally, the fitted model is such that $f_i(\cdot;\theta)$ is a Gaussian distribution with mean $1000/|\theta - r_i|^2$ and variance σ_*^2 (see [15] for a similar model).

The convergence of the algorithm (2) to the consensus subspace is illustrated in Figure 2. Four paths (started from the same value $\boldsymbol{\theta}_0$) are run and we display $n \mapsto (1/N)|\boldsymbol{\theta}_n - \mathbb{1} \otimes \theta_*|$ for $n \leq 50000$. Convergence to the consensus subspace can be observed; note also the role of the step size sequence in the rate of convergence (compare the definition of γ_n above and the changes in the slopes at time $n = 10000$ and $n = 20000$).

The rate in the Central Limit Theorem is illustrated in Figure 3. We compute $(1/N)\gamma_n^{-1}\mathcal{E}_n$ where $\mathcal{E}_n := |\boldsymbol{\theta}_n - \mathbb{1} \otimes \theta_*|^2$ for 500 independent paths of the algorithm (2) started from the same value $\boldsymbol{\theta}_0$. Figure 3 shows the median, and the first and third quartiles of these 500 values as a function of the number of iterations n.

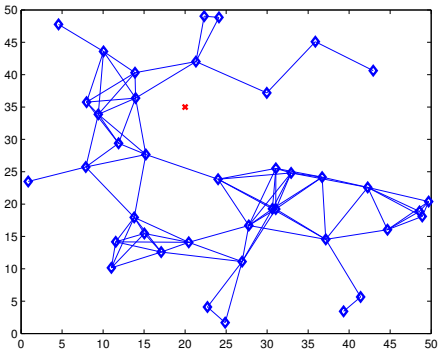

Figure 1: $N = 40$ sensors (diamonds), their neighborhood (lines) and the source (red star)

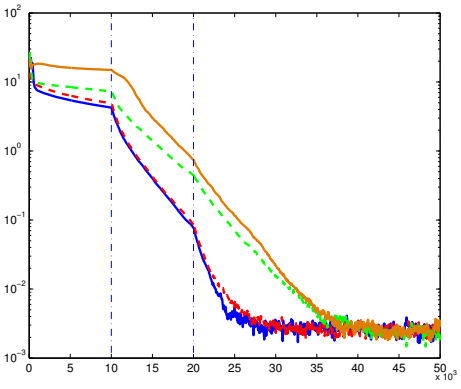

Figure 2: Cumulated relative error (over the N sensors) when estimating θ_*, as a function of the number of iterations.

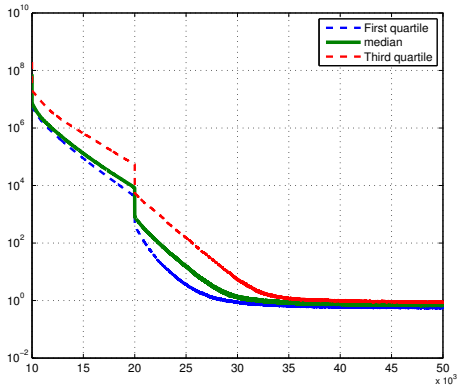

Figure 3: First quartile, median and third quartile of the distribution of $\gamma_n^{-1}\mathcal{E}_n$, as a function of the number of iterations n, estimated from 500 Monte Carlo runs.

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] 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.

[4] F.P. Kelly, A.K. Maulloo, and D.K.H. Tan, "Rate control for communication networks: shadow prices, proportional fairness and stability," *Journal of the Operational Research society*, vol. 49, no. 3, pp. 237–252, 1998.

[5] 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.

[6] 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*, vol. 56, no. 1, pp. 350–364, 2008.

[7] S. Barbarossa and G. Scutari, "Decentralized maximum-likelihood estimation for sensor networks composed of nonlinearly coupled dynamical systems," *IEEE Trans. on Signal Processing*, vol. 55, no. 7, pp. 3456–3470, July 2007.

[8] D.P. Bertsekas and J.N. Tsitsiklis, *Parallel and distributed computation*, Prentice Hall Inc., 1989.

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

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

[15] M. Rabbat and R. Nowak, "Distributed optimization in sensor networks," in *Proceedings of the 3rd international symposium on Information processing in sensor networks*. ACM, 2004, pp. 20–27.