Performance of a Distributed Stochastic Approximation Algorithm

Pascal Bianchi, *Member, IEEE*, Gersende Fort, and Walid Hachem, *Member, IEEE*

***Abstract*—In this paper, a distributed stochastic approximation algorithm is studied. Applications of such algorithms include decentralized estimation, optimization, control or computing. The algorithm consists in two steps: a local step, where each node in a network updates a local estimate using a stochastic approximation algorithm with decreasing step size, and a gossip step, where a node computes a local weighted average between its estimates and those of its neighbors. Convergence of the estimates toward a consensus is established under weak assumptions. The approach relies on two main ingredients: the existence of a Lyapunov function for the mean field in the agreement subspace, and a contraction property of the random matrices of weights in the subspace orthogonal to the agreement subspace. A second-order analysis of the algorithm is also performed under the form of a central limit Theorem. The Polyak-averaged version of the algorithm is also considered.**

***Index Terms*—Convergence, decentralized estimation, decentralized optimization, gossip algorithms, stochastic approximation.**

I. Introduction

STOCHASTIC approximation has been a very active research area for the last 60 years (see e.g., [1], [2]). The pattern for a stochastic approximation algorithm is provided by the recursion $\theta_n = \theta_{n-1} + \gamma_n Y_n$, where θ_n is typically a $\mathbb{R}^d$-valued sequence of parameters, Y_n is a sequence of random observations, and γ_n is a deterministic sequence of step sizes. An archetypal example of such algorithms is provided by stochastic gradient algorithms. These are characterized by the fact that $Y_n = -\nabla g(\theta_{n-1}) + \xi_n$ where ∇g is the gradient of a function g to be minimized, and where $(\xi_n)_{n\geq 0}$ is a noise sequence corrupting the observations.

In the traditional setting, sensing, and processing capabilities needed for the implementation of a stochastic approximation algorithm are centralized on one machine. Alternatively, distributed versions of these algorithms where the updates are done by a network of communicating nodes (or agents) have recently aroused a great deal of interest. Applications include decentralized estimation, control, optimization, and parallel computing.

In this paper, we consider a network composed by N nodes (sensors, robots, computing units, and so on). Node i generates a $\mathbb{R}^d$-valued stochastic process $(\theta_{n,i})_{n\geq 1}$ through a two-step iterative algorithm: a local and a so-called gossip step. At time n:

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

$$\tilde{\theta}_{n,i} = \theta_{n-1,i} + \gamma_n Y_{n,i}, \tag{1}$$

where γ_n is a deterministic positive step size and where the $\mathbb{R}^d$-valued random process $(Y_{n,i})_{n\geq 1}$ represents the observations made by agent i.

[`Gossip step`] Node i is able to observe the values $\tilde{\theta}_{n,j}$ of some other js and computes the weighted average

$$\theta_{n,i} = \sum_{j=1}^{N} w_n(i,j)\, \tilde{\theta}_{n,j}, \tag{2}$$

where the $w_n(i,j)$s are scalar nonnegative random coefficients such that $\sum_{j=1}^{N} w_n(i,j) = 1$ for any i. The sequence of random matrices $W_n := [w_n(i,j)]_{i,j=1}^{N}$ represents the time-varying communication network between the nodes.

Contributions. This paper studies a distributed stochastic approximation algorithm in the context of random row-stochastic gossip matrices W_n.

1) Under the assumption that the algorithm is stable, we prove convergence of the algorithm to the sought consensus. The unanimous convergence of the estimates is also established in the case where the frequency of information exchange between the nodes converges to zero at some controlled rate. In practice, this means that matrices W_n become more and more likely to be equal to identity as $n \to \infty$. The benefits of this possibility in terms of power devoted to communications are obvious.
2) We provide verifiable sufficient conditions for stability.
3) We establish a central limit theorem (CLT) on the estimates in the case where the W_n are doubly stochastic. We show in particular that the node estimates tend to fluctuate synchronously for large n, i.e., the disagreement between the nodes is negligible at the CLT scale. Interestingly, the distributed algorithm under study has the same asymptotic variance as its centralized analog.
4) We also consider a CLT on the sequences averaged over time as introduced in [3]. We show that averaging always improves the rate of convergence and the asymptotic variance.

Motivations and examples. The algorithm under study is motivated by the emergence of various decentralized network structures such as sensor networks, computer clouds or wireless ad hoc networks. One of the main application targets is distributed optimization. In this context, one seeks to minimize

Manuscript received March 07, 2012; revised December 22, 2012; accepted June 28, 2013. Date of publication August 02, 2013; date of current version October 16, 2013. This work was supported in part by the French National Research Agency under the program ANR-07 ROBO 002.

The authors are with the LTCI—CNRS/Telecom ParisTech, 75634 Paris Cedex 13, France (e-mail: bianchi@telecom-paristech.fr; gersende.fort@telecom-paristech.fr; walid.hachem@telecom-paristech.fr).

Communicated by G. V. Moustakides, Associate Editor for Detection and Estimation.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2013.2275131

a sum of some local objective differentiable functions f_i of the agents

$$\texttt{Minimize} \sum_{i=1}^{N} F_i(\theta). \tag{3}$$

Function F_i is supposed to be unknown by any other agent $j \neq i$. In this context, the distributed algorithm (1)–(2) would reduce to a distributed stochastic gradient algorithm by letting $Y_{n,i} = -\nabla_\theta F_i(\theta_{n-1,i}) + \xi_{n,i}$ where ∇_θ is the gradient w.r.t. θ and $\xi_{n,i}$ represents some possible random perturbation $\xi_{n,i}$ at time n.

In a machine learning context, F_i is typically the risk function of a classifier indexed by θ and evaluated based on a local training set at agent i [4]. In a wireless ad-hoc network, F_i represents some (negative) performance measure of a transmission such as the Shannon capacity, and the aim is typically to search for a relevant resource allocation vector θ (see [5] for more details). As a third example, an application framework to statistical estimation is provided in Section V. In that case, it is assumed that node i receives some i.i.d. time series $(X_{n,i})_n$ with probability density function $f_*(x)$. The system designer considers that the density of $(X_{n,1}, \ldots, X_{n,N})$ belongs to a parametric family $\{f(\theta, \boldsymbol{x})\}_\theta$ where $f(\theta, \boldsymbol{x}) = \prod_{i=1}^{n} f_i(\theta, x_i)$. Then, a well-known contrast for the estimation of θ is given by the Kullback–Leibler divergence $D(f_* \,\|\, f(\theta, \cdot))$ [6]. Finding a minimizer boils down to the minimization of (3) by setting $F_i(\theta) = D(f_{i,*} \,\|\, f_i(\theta, \cdot))$ where $f_{i,*}$ is the ith marginal of f_*. Then, algorithm (1)–(2) coincides with a distributed online maximum likelihood (ML) estimator by setting $Y_{n,i} = -\nabla_\theta \log f_i(\theta_{n-1,i}, X_{n,i})$. Under some regularity conditions, it can be easily checked that $Y_{n,i} = -\nabla_\theta F_i(\theta_{n-1,i}) + \xi_{n,i}$ where $\xi_{n,i}$ is a martingale increment sequence.

Position w.r.t. existing works. There is a rich literature on distributed estimation and optimization algorithms, see [7]–[13] as a nonexhaustive list. Among the first gossip algorithms are those considered in the treatise [14] and in [15], as well as in [16], the latter reference dealing with the case of a constant step size. The case where the gossip matrices are random and the observations are noiseless is considered in [17]. Nedic *et al.* [11] solve a constrained optimization by also using noiseless estimates. The contributions [10] and [13] consider the framework of linear regression models.

In this paper, the random gossip matrices W_n are assumed to be row stochastic, i.e., $W_n \mathbf{1} = \mathbf{1}$ where $\mathbf{1}$ is the vector whose components equal one, and column stochastic in the mean, i.e., $\mathbf{1}^T \mathbb{E}[W_n] = \mathbf{1}^T$. Observe that the row stochasticity constraint $W_n \mathbf{1} = \mathbf{1}$ is local, since it simply requires that each agent makes a weighted sum of the estimates of its neighbors with weights summing to one. Alternatively, the column stochasticity constraint $\mathbf{1}^T W_n = \mathbf{1}^T$ which is assumed in many contributions (see e.g., [18], [11], [19], [20]) requires a coordination at the network level (nodes must coordinate their weights). This constraint is not satisfied by a large class of gossip algorithms. As an example, the well-known broadcast gossip matrices (see Section II-B) are only column stochastic in the mean. As opposed to the aforementioned papers, it is worth noting that some works such as [16], [12], [5] get rid of the column-stochasticity condition. As a matter of fact, assumption $\mathbf{1}^T \mathbb{E}[W_n] = \mathbf{1}^T$ is even relaxed in [16]. Nevertheless, considering for instance Problem (3), this comes at the price of losing the convergence to the sought minima.

In many contributions (see e.g., [16], [8], or [10]), the gossip step is performed before the local step, contrary to what is done in this paper. The general techniques used in this paper to establish the convergence toward a consensus, the stability and the fluctuations of the estimates can be adapted without major difficulty to that situation.

In [19], projected stochastic (sub)gradient algorithms are considered in the case where matrices $(W_n)_n$ are doubly stochastic. Such results have later been extended by [5] to the case of nonconvex optimization, also relaxing the doubly stochastic assumption. It is worth noting that such works explicitly or implicitly rely on a projection step onto a compact convex set. In many scenarios (such as unconstrained optimization for example), the estimate is not naturally supposed to be confined into a known compact set. In that case, introducing an artificial projection step is known to modify the limit points of the algorithm. On the opposite, this paper addresses the issue of unprojected stochastic approximation algorithms. In this context, stability turns out to be a crucial issue which is addressed in this paper. Note that the stability issues are not considered in most of [16]. Finally, unlike previous works such as [19] or [5], we also address the issue of convergence rate and characterize the asymptotic fluctuations of the estimation error.

From a methodological viewpoint, our analysis does not rely on convex optimization tools such as in e.g., [18], [11], [19]) and does not rely on perturbed differential inclusions as in [5]. The almost sure convergence result is obtained following an approach inspired by [21] and [22] (other works such as [16] consider weak convergence approaches). The stability result is obtained by introducing a Lyapunov function and by jointly controlling the moments of this Lyapunov function and the second order moments of the disagreements between local estimates. Finally, the study of the asymptotic fluctuations of the estimate is based on recent results of [23] and is partly inspired by the works of [24].

This paper is organized as follows. In Section II, we state and comment our basic assumptions. The algorithm convergence is studied in Section III. The second-order behavior of the algorithm is described in Section IV. An application relative to distributed estimation is described in Section V, along with some numerical simulations. The appendix is devoted to the proofs.

II. Model and the Basic Assumptions

Let us start by writing the distributed algorithm described in the previous section in a more compact form. Define the $\mathbb{R}^{dN}$-valued random vectors $\boldsymbol{\theta}_n$ and $\boldsymbol{Y}_n$ by $\boldsymbol{\theta}_n := (\theta_{n,1}^T, \ldots, \theta_{n,N}^T)^T$ and $\boldsymbol{Y}_n := (Y_{n,1}^T, \ldots, Y_{n,N}^T)^T$ where A^T denotes the transpose of the matrix A. The algorithm reduces to

$$\boldsymbol{\theta}_n = (W_n \otimes I_d)\,(\boldsymbol{\theta}_{n-1} + \gamma_n \boldsymbol{Y}_n)\,, \tag{4}$$

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

Note that we always assume $\mathbb{E}|\boldsymbol{\theta}_0|^2 < \infty$ throughout the paper, where $|\,.\,|$ represents the Euclidean norm.

Remark 1: Following [3], we also consider the averaged sequence $(\bar{\boldsymbol{\theta}}_n)_{n\geq 1}$, where $\bar{\boldsymbol{\theta}}_n := (\bar{\theta}_{n,1}^T, \ldots, \bar{\theta}_{n,N}^T)^T$ and the components are given by

$$\bar{\theta}_{n,i} = \frac{1}{n}\sum_{k=1}^{n} \theta_{k,i} \tag{5}$$

at any instant n for node i. We will show in Section IV-B that this averaging technique improves the convergence rate of the distributed stochastic approximation algorithm. In this paper, we analyze the asymptotic behavior of both sequences $\bar{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}_n$ as $n \to \infty$.

A. Observation and Network Models

Let $(\mu_{\boldsymbol{\theta}})_{\boldsymbol{\theta}\in\mathbb{R}^{dN}}$ be a family of probability measures on $\mathbb{R}^{dN}$ endowed with its Borel σ-field $\mathcal{B}(\mathbb{R}^{dN})$ such that for any $A \in \mathcal{B}(\mathbb{R}^{dN})$, $\boldsymbol{\theta} \mapsto \mu_{\boldsymbol{\theta}}(A)$ is measurable from $\mathcal{B}(\mathbb{R}^{dN})$ to $\mathcal{B}([0,1])$ where $\mathcal{B}([0,1])$ denotes the Borel σ-field on $[0,1]$.

We consider the case when the random process $(\boldsymbol{Y}_n, W_n)_{n\geq 1}$ is adapted to a filtered probability space $(\Omega, \mathcal{A}, \mathbb{P}, (\mathcal{F}_n)_{n\geq 0})$ and satisfy

Assumption 1:

a) $(W_n)_{n\geq 1}$ is a sequence of $N \times N$ random matrices with nonnegative elements such that
   i) W_n is row stochastic: $W_n\mathbf{1} = \mathbf{1}$,
   ii) $\mathbb{E}(W_n)$ is column stochastic: $\mathbf{1}^T\mathbb{E}(W_n) = \mathbf{1}^T$,

b) For any positive measurable functions f, g, and any $n \geq 0$,

$$\mathbb{E}[f(W_{n+1})g(\boldsymbol{Y}_{n+1})|\mathcal{F}_n] = \mathbb{E}[f(W_{n+1})] \int g(\boldsymbol{y})\mu_{\boldsymbol{\theta}_n}(d\boldsymbol{y}). \tag{6}$$

c) The sequence $(W_n)_{n\geq 1}$ is identically distributed and the spectral norm ρ of matrix $\mathbb{E}(W_1^T(I_N - \mathbf{1}\mathbf{1}^T/N)W_1)$ satisfies $\rho < 1$.

Assumptions 1a and 1c capture the properties of the gossiping scheme within the network. Following the work of [17], random gossip is assumed in this paper. Assumption 1a has been commented in Section I. The assumption on the spectral norm in Assumption 1c is a connectivity condition of the underlying network graph which will be discussed in more details in Section II-B. Assumption 1b implies that 1) the random variables (r.v.) W_n and $\boldsymbol{Y}_n$ are independent conditionally to the past, 2) the r.v. $(W_n)_{n\geq 1}$ are independent, and 3) the conditional distribution of $\boldsymbol{Y}_{n+1}$ given the past is $\mu_{\boldsymbol{\theta}_n}$. This assumption is quite usual in the framework of stochastic approximation and is sometimes refer to as a Robbins–Monro setting. As a particular case, this assumption holds if $\boldsymbol{Y}_{n+1}$ has the form $\boldsymbol{Y}_{n+1} = g(\boldsymbol{\theta}_n) + \xi_{n+1}$ where ξ_{n+1} is an i.i.d. process.

It is also assumed that the step-size sequence $(\gamma_n)_{n\geq 1}$ in the stochastic approximation scheme (1) satisfies the following conditions which are rather usual in the framework of stochastic approximation algorithms [2].

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

B. Illustration: Some Examples of Gossip Schemes

We describe three standard gossip schemes so-called *pairwise*, *broadcast*, and *dropout* schemes. The reader may refer to [25] for a more complete picture and for more general gossip strategies. The network of agents is represented as a nondirected graph (E, V) where E is the set of edges and V is the set of N vertices.

1) Pairwise Gossip: This example can be found in [17] on average consensus (see also [5]).

At time n, two connected nodes—say i and j—wake up, independently from the past. Nodes i and j compute the weighted average $\theta_{n,i} = \theta_{n,j} = 0.5\tilde{\theta}_{n,i} + 0.5\tilde{\theta}_{n,j}$; and for $k \notin \{i,j\}$, the nodes do not gossip: $\theta_{n,k} = \tilde{\theta}_{n,k}$. In this example, given the edge $\{i,j\}$ wakes up, W_n is equal to $I_N - (e_i - e_j)(e_i - e_j)^T/2$ where e_j denotes the ith vector of the canonical basis in $\mathbb{R}^N$; and the matrices $(W_n)_{n\geq 0}$ are i.i.d. and doubly stochastic. Assumption 1a is obviously satisfied. Conditions for Assumption 1c can be found in [17]: the spectral norm ρ of the matrix $\mathbb{E}(W_n(I_N - \mathbf{1}\mathbf{1}^T/N)W_n^T)$ is in $[0,1)$ if and only if the weighted graph $(\mathsf{E}, \mathsf{V}, \mathsf{W})$ is connected, where the wedge $\{i,j\}$ is weighted by the probability that the nodes i, j communicate.

2) Broadcast Gossip: This example is adapted from the broadcast scheme in [26]. At time n, a node i wakes up at random with uniform probability and broadcasts its temporary update $\tilde{\theta}_{n,i}$ to all its neighbors $\mathcal{N}_i$. Any neighbor j computes the weighted average $\theta_{n,j} = \beta\tilde{\theta}_{n,i} + (1-\beta)\tilde{\theta}_{n,j}$. On the other hand, the nodes k which do not belong to the neighborhood of i (including i itself) sets $\theta_{n,k} = \tilde{\theta}_{n,k}$. Note that, as opposed to the pairwise scheme, the transmitter node i does not expect any feedback from its neighbors. Then, given i wakes up, the (k,ℓ)th component of W_n is given by

$$w_n(k,\ell) = \begin{cases} 1 & \text{if } k \notin \mathcal{N}_i \text{ and } k = \ell, \\ \beta & \text{if } k \in \mathcal{N}_i \text{ and } \ell = i, \\ 1-\beta & \text{if } k \in \mathcal{N}_i \text{ and } k = \ell, \\ 0 & \text{otherwise.} \end{cases}$$

This matrix W_n is not doubly stochastic but $\mathbf{1}^T\mathbb{E}(W_n) = \mathbf{1}^T$ (see for instance [26]). Thus, the matrices $(W_n)_{n\geq 1}$ are i.i.d. and satisfy the Assumption 1a. Here again, it can be shown that the spectral norm ρ of $\mathbb{E}(W_n(I_N - \mathbf{1}\mathbf{1}^T/N)W_n^T)$ is in $[0,1)$ if and only if (E, V) is a connected graph (see [26]).

3) Network Dropouts: In this simple example, the network is subjected from time to time to a dropout: consider any sequence of gossip matrices W_n satisfying Assumptions 1a and 1c, and put $W'_n = B_nW_n + (1 - B_n)I_N$ where B_n is a sequence of i.i.d. Bernoulli random variables independent of the W_n. The network whose gossip matrices are the W'_n incurs a dropout at the moments where $B_n = 0$. At these moments, the nodes locally update their estimates and skip the gossip step. It is easy to show that the sequence W'_n satisfies Assumptions 1a and 1c.

III. Convergence Results

In this section, we address the asymptotic behavior when $n \to \infty$ of the algorithm (4) and of its averaged version (5). To that goal, we write $\boldsymbol{\theta}_n$ as the sum of a vector in the consensus space and a disagreement vector. Let

$$J := (\mathbf{1}\mathbf{1}^T/N) \otimes I_d, \qquad J_\perp := I_{dN} - J, \tag{7}$$

be resp. the projector onto the *consensus subspace* $\{\mathbf{1}\otimes\theta : \theta\in\mathbb{R}^d\}$ and the projector onto the orthogonal subspace. For any vector $\boldsymbol{x}\in\mathbb{R}^{dN}$, define the vector of $\mathbb{R}^d$

$$\langle\boldsymbol{x}\rangle := \frac{1}{N}(\mathbf{1}^T\otimes I_d)\boldsymbol{x}, \tag{8}$$

so that $J\boldsymbol{x}=\mathbf{1}\otimes\langle\boldsymbol{x}\rangle$. Note that $\langle\boldsymbol{x}\rangle=(x_1+\cdots+x_N)/N$ in case we write $\boldsymbol{x}=(x_1^T,\ldots,x_N^T)^T$, x_i in $\mathbb{R}^d$. Set

$$\boldsymbol{x}_\perp := J_\perp\boldsymbol{x} \tag{9}$$

so that $\boldsymbol{x}=\mathbf{1}\otimes\langle\boldsymbol{x}\rangle+\boldsymbol{x}_\perp$. We will refer to $\boldsymbol{\theta}_{\perp,n}:=J_\perp\boldsymbol{\theta}_n$ as the *disagreement vector*.

The convergence results rely on the following equations: under Assumption 1a, it holds

$$\langle\boldsymbol{\theta}_n\rangle=\langle\boldsymbol{\theta}_{n-1}\rangle+\gamma_n\langle(W_n\otimes I_d)(\boldsymbol{Y}_n+\gamma_n^{-1}\boldsymbol{\theta}_{\perp,n-1})\rangle, \tag{10}$$

$$\gamma_{n+1}^{-1}\boldsymbol{\theta}_{\perp,n}=\frac{\gamma_n}{\gamma_{n+1}}J_\perp(W_n\otimes I_d)\left(\gamma_n^{-1}\boldsymbol{\theta}_{\perp,n-1}+J_\perp\boldsymbol{Y}_n\right). \tag{11}$$

We then first address the almost sure convergence of the sequence $(\boldsymbol{\theta}_n)_{n\geq1}$ 1) by showing that the nonhomogeneous controlled Markov chain $(\gamma_{n-1}^{-1}\boldsymbol{\theta}_{\perp,n})_n$ is stable enough so that $(\boldsymbol{\theta}_{\perp,n})_n$ converges almost surely to zero and, 2) by applying results on the convergence of stochastic approximation algorithms with state-dependent noise in order to identify the limiting points of the sequence $(\langle\boldsymbol{\theta}_n\rangle)_{n\geq1}$. These results are stated in Theorem 1 (and Theorem 2 in the case of vanishing communication rate); we prove that all agents eventually reach an agreement on the value of their estimate: the limit points of $(\boldsymbol{\theta}_n)_{n\geq1}$ (resp. $(\bar{\boldsymbol{\theta}}_n)_{n\geq1}$) given by (4) (resp. (5)) are of the form $\mathbf{1}\otimes\theta_\star$.

It is known that convergence of stochastic approximation algorithms to an attractive set is established provided that the sequence remains in a compact set with probability one and is, with probability 1, infinitely often in the domain of attraction of this attractive set. Our convergence result is stated under assumptions implying the recurrence property provided the sequence remains almost-surely in a compact set. Therefore, our convergence results are derived under a boundedness assumption, and we then provide in Theorem 3 sufficient conditions for this boundedness condition to be satisfied.

All these convergence results are obtained under conditions on the state-dependent noise sequence in the stochastic approximation scheme (10). These conditions roughly speaking assume 1) that there exist a Lyapunov function and an attractive set associated with the mean field of the noisy ordinary differential (10), 2) regularity-in-$\boldsymbol{\theta}$ of the probability distributions $(\mu_{\boldsymbol{\theta}})_{\boldsymbol{\theta}\in\mathbb{R}^{dN}}$. The exact assumptions are stated herein.

A. Assumptions on the Distributions $\mu_{\boldsymbol{\theta}}$

Define the function $h:\mathbb{R}^d\to\mathbb{R}^d$ by

$$h(\theta):=\int\langle\boldsymbol{y}\rangle\,\mu_{\mathbf{1}\otimes\theta}(d\boldsymbol{y}). \tag{12}$$

We shall refer to h as the *mean field*. The key ingredient to prove the convergence of a stochastic approximation procedure is the existence of a Lyapunov function V for the mean field h, i.e., a function $V:\mathbb{R}^d\to\mathbb{R}^+$ such that $\nabla V^T\,h\leq0$. Precisely, it is assumed.

Assumption 3: There exists a function $V:\mathbb{R}^d\to\mathbb{R}^+$ such that

a) V is continuously differentiable.
b) For any $\theta\in\mathbb{R}^d$, $\nabla V(\theta)^Th(\theta)\leq0$, where h is given by (12).
c) For any $M>0$, the level set $\{\theta\in\mathbb{R}^d : V(\theta)\leq M\}$ is compact.
d) The set $\mathcal{L}:=\{\theta\in\mathbb{R}^d:\nabla V(\theta)^Th(\theta)=0\}$ is nonempty and there exists M_0 such that $\mathcal{L}\subseteq\{V\leq M_0\}$.
e) The function h given by (12) is continuous on $\mathbb{R}^d$.
f) $V(\mathcal{L}):=\{V(\theta):\theta\in\mathcal{L}\}$ has an empty interior.

Observe that Assumptions 3d and 3f are trivially satisfied when $\mathcal{L}$ is finite.

When h is a gradient field i.e., $h=-\nabla g$, a natural candidate for the Lyapunov function is $V=g$. In this case, $\mathcal{L}=\{\nabla g=0\}$; when g is d-times differentiable, Sard's theorem implies that $g(\{\nabla g=0\})$ has an empty interior. If g is strictly convex and it reaches its minimum at a finite $\theta_\star$, the function $\theta\mapsto|\theta-\theta_\star|^2$ is also a Lyapunov function. In this case, $\mathcal{L}=\{\theta_\star\}$.

Assumption 4: For any $M>0$,

a) $\sup_{|\boldsymbol{\theta}|\leq M}\int|\boldsymbol{y}|^2\,\mu_{\boldsymbol{\theta}}(d\boldsymbol{y})<\infty$.
b) there exists a constant C_M such that for any $|\boldsymbol{\theta}|\leq M$,

$$\left|\int\langle\boldsymbol{y}\rangle\,\mu_{\boldsymbol{\theta}}(d\boldsymbol{y})-\int\langle\boldsymbol{y}\rangle\,\mu_{\mathbf{1}\otimes\langle\boldsymbol{\theta}\rangle}(d\boldsymbol{y})\right|\leq C_M|\boldsymbol{\theta}_\perp|. \tag{13}$$

The condition (13) is a regularity condition on the distribution of $\langle\boldsymbol{Y}_{n+1}\rangle$ given the past.

B. Almost Sure Convergence of the Distributed Algorithm

Define $\mathsf{d}(\theta,A):=\inf\{|\theta-\varphi|:\varphi\in A\}$ for any $\theta\in\mathbb{R}^d$ and $A\subset\mathbb{R}^d$.

Theorem 1: Let us consider Assumptions 1–4. Assume in addition that $\lim_n\gamma_n/\gamma_{n-1}=1$ and

$$\mathbb{P}\left\{\limsup_n|\boldsymbol{\theta}_n|<\infty\right\}=1. \tag{14}$$

Then, with probability 1

$$\lim_{n\to\infty}\mathsf{d}(\langle\boldsymbol{\theta}_n\rangle,\mathcal{L})=0,\qquad\lim_n\boldsymbol{\theta}_{\perp,n}=0, \tag{15}$$

where $\mathcal{L}$ is given by Assumption 3. Moreover, with probability one, $(\langle\boldsymbol{\theta}_n\rangle)_{n\geq1}$ converges to a connected component of $\mathcal{L}$.

Theorem 1 is proved in Appendix B. Theorem 1 shows that when the stability condition (14) holds true, the vector of iterates $\boldsymbol{\theta}_n$ given by (4) converges almost surely to the consensus space as $n\to\infty$ so that the network asymptotically achieves consensus. Moreover, this consensus belongs to the attractive set of the Lyapunov function.

Since V is continuous, Theorem 1 implies that with probability 1 (w.p.1), the sequence $\{V(\langle\boldsymbol{\theta}_n\rangle)\}_{n\geq0}$ converges to a (random) point $\upsilon_\star\in V(\mathcal{L})$. This can be used to show that $(\langle\boldsymbol{\theta}_n\rangle)_{n\geq0}$ converges to a connected component of $\{\theta\in\mathcal{L}:V(\theta)=\upsilon_\star\}$. In general, this does not imply that $(\langle\boldsymbol{\theta}_n\rangle)_{n\geq0}$ converges w.p.1 to some (random point) $\theta_\star\in\mathcal{L}$. Note nevertheless that this holds true w.p.1 when $\mathcal{L}$ is finite.

Along any sequence $(\boldsymbol{\theta}_n)_{n\geq0}$ converging to $\mathbf{1}\otimes\theta_\star$ for some $\theta_\star\in\mathcal{L}$, Cesaro's lemma implies that the averaged sequence

$(\bar{\boldsymbol{\theta}}_n)_{n\geq 0}$ converges w.p.1 to $\mathbf{1}\otimes\theta_\star$. Therefore, the averaged sequence (5) and the original sequence (4) have the same limiting value, if any.

C. Case of a Vanishing Communication Rate

Theorems 1 still holds true when the r.v. $(W_n)_{n\geq 1}$ are not identically distributed. An interesting example is when $\mathbb{P}\{W_n = I_N\} \to 1$ as $n\to\infty$. From a communication point of view, this means that the exchange of information between agents becomes rare as $n\to\infty$. This context is especially interesting in case of wireless networks, where it is often required to limit as much as possible the amount of communication between the nodes.

In such cases, Assumption 1c does no longer hold true. We prove a convergence result for the algorithms (4) and (5) when the spectral norm of the matrix $\mathbb{E}(W_n^T(I_N - \mathbf{1}\mathbf{1}^T/N)W_n)$ and the step size sequence $(\gamma_n)_{n\geq 1}$ satisfy the following assumption.

Assumption 5: $\sum_n \gamma_n = \infty$ and there exists $\alpha > 1/2$ such that

$$\lim_{n\to\infty} n^\alpha\gamma_n = 0, \qquad \lim_{n\to\infty} n^{1+\alpha}\gamma_n = +\infty, \quad (16)$$

$$\liminf_{n\to\infty} \frac{1-\rho_n}{n^\alpha\gamma_n} > 0, \quad (17)$$

where ρ_n is the spectral norm of the matrix $\mathbb{E}(W_n^T(I_N - \mathbf{1}\mathbf{1}^T/N)W_n)$.

Note that under Assumption 5, $\lim_n n(1-\rho_n) = +\infty$. A typical framework where this assumption is useful is the following. Let $(B_n)_n$ be a Bernoulli sequence of independent r.v. with $\mathbb{P}(B_n = 1) = p_n$ and the probabilities p_n decrease in such a way that $\liminf_n p_n/(n^\alpha\gamma_n) > 0$: replace the matrices W_n described by Assumption 1 with $B_nW_n + (1-B_n)I_N$. Here p_n represents the probability that a communication between the nodes takes place at time n.

We also have $\sum_n \gamma_n^2 < \infty$ so that the step-size sequence $(\gamma_n)_{n\geq 1}$ satisfies the standard conditions for stochastic approximation scheme to converge.

An example of sequences $(\gamma_n)_{n\geq 1}$, $(\rho_n)_{n\geq 1}$ satisfying Assumption 5 is given by $1-\rho_n = a/n^\eta$ and $\gamma_n = \gamma_0/n^\xi$ with η, ξ such that $0\leq\eta<\xi-1/2\leq 1/2$. In particular, $\xi\in(1/2,1]$ and $\eta\in[0,1/2)$.

When the r.v. $(W_n)_{n\geq 1}$ are i.i.d., the spectral norm ρ_n is equal to ρ for any n, and (17) implies $\rho<1$: one is back to Assumption 1c. From this point of view, Assumption 5 is weaker than Assumption 1c. Nevertheless, stronger constraints than Assumption 1c are needed on the step size $(\gamma_n)_{n\geq 1}$.

When substituting Assumption 1c by Assumption 5, we have following theorem.

Theorem 2: The statement of Theorem 1 remains valid under Assumptions 1a, 1b, and 2–5 and (14).

Theorem 2 is proved in Appendix B.

D. Stability

In this section, we provide sufficient conditions implying (14). These conditions are stated in the case of a vanishing communication rate but remain valid when Assumption 5 is replaced with Assumption 1c. The proof of Theorem 3 is given in Appendix C.

Theorem 3: Let us consider Assumptions 1a, 1b, 2, 3a–3e, and 5. Assume in addition that

ST1. ∇V is Lipschitz on $\mathbb{R}^d$.

ST2. there exists a constant C such that for any $\boldsymbol{\theta}\in\mathbb{R}^{dN}$,

$$\int |\boldsymbol{y}|^2\,\mu_{\boldsymbol{\theta}}(d\boldsymbol{y}) \leq C\left(1+V(\langle\boldsymbol{\theta}\rangle)+|\boldsymbol{\theta}_\perp|^2\right),$$

$$\left|\int\langle\boldsymbol{y}\rangle\,\mu_{\boldsymbol{\theta}}(d\boldsymbol{y}) - \int\langle\boldsymbol{y}\rangle\,\mu_{\mathbf{1}\otimes\langle\boldsymbol{\theta}\rangle}(d\boldsymbol{y})\right| \leq C|\boldsymbol{\theta}_\perp|.$$

Then, $\mathbb{P}\{\limsup_n|\boldsymbol{\theta}_n|<\infty\}=1$.

It is proved in Appendix C that under the assumptions of Theorem 3, a stronger result holds (see Lemma 5): the sequence $(\boldsymbol{\theta}_{\perp,n})_{n\geq 1}$ converges to zero with probability 1 and $(\langle\boldsymbol{\theta}_n\rangle)_{n\geq 1}$ is stable in the sense that $\sup_n V(\langle\boldsymbol{\theta}_n\rangle)<\infty$.

Note that the Lipschitz assumption on the gradient ∇V combined with Assumption ST2 implies that h is at most linearly increasing when $|\boldsymbol{\theta}|\to\infty$.

The stability condition (14) could also be satisfied by modifying the algorithm (4) with a truncation step. Truncation on a fixed compact set of $\mathbb{R}^{dN}$ is easy to implement and natural when constraints on the system are available *a priori*; nevertheless it becomes impractical and questionable in many situations of interest when a compact set containing the limiting set $\mathcal{L}$ is not known *a priori*. Another stability strategy consists in truncations on randomly varying compact sets [27]; derivation of conditions implying the stability of Algorithm (4) without modifying its limiting set under such an approach is out of the scope of this paper and left to the interested reader.

IV. Convergence Rates

In this section, we derive the convergence rate in L^2 of the disagreement sequence $(\boldsymbol{\theta}_{\perp,n})_n$ defined $\boldsymbol{\theta}_{\perp,n} := J_\perp\boldsymbol{\theta}_n$ [see (7) and (9)]. We also derive central limit theorems for the sequences $(\boldsymbol{\theta}_n)_n$ and $(\bar{\boldsymbol{\theta}}_n)_n$: we show that averaging always improves the convergence rate and the asymptotic variance.

A. Convergence Rate of the Disagreement Vector $\boldsymbol{\theta}_{\perp,n}$

Whereas Theorem 1 states that $\lim_n\boldsymbol{\theta}_{\perp,n}=0$ almost surely, Theorem 4 provides an information on the convergence rate: $\boldsymbol{\theta}_{\perp,n}$ tends to zero in L^2 at rate $1/\gamma_n$. For a positive deterministic sequence $(a_n)_{n\geq 1}$, $\mathcal{O}(a_n)$ stands for a deterministic $\mathbb{R}^\ell$-valued sequence $(x_n)_{n\geq 1}$ such that $\sup_n a_n^{-1}|x_n|<\infty$. The proof of Theorem 4 is given in Appendix D.

Theorem 4: Let us consider Assumptions 1, 2, and 4a. For any $M>0$,

$$\gamma_n^{-2}\mathbb{E}\left(|\boldsymbol{\theta}_{\perp,n}|^2\mathbf{1}_{\sup_{k\leq n-1}|\boldsymbol{\theta}_k|\leq M}\right) \leq \frac{\rho\,C}{(1-\sqrt{\rho})^2} + \mathcal{O}\left(\rho^n\gamma_n^{-2}\right) \quad (18)$$

where ρ is given by Assumption 1c and where $C := \limsup_{n\to\infty}\mathbb{E}(|\boldsymbol{Y}_{\perp,n}|^2\mathbf{1}_{\sup_{k\leq n-1}|\boldsymbol{\theta}_k|\leq M})$ is finite.

B. Central Limit Theorems

We derive central limit theorems for sequences $(\boldsymbol{\theta}_n)_n$ and $(\bar{\boldsymbol{\theta}}_n)_n$ converging to a point $\mathbf{1}\otimes\theta_\star$ for some $\theta_\star\in\mathcal{L}$. To that goal, we restrict our attention to the case when the matrix $(W_n)_n$ are doubly stochastic, i.e., $\mathbf{1}^TW_n = \mathbf{1}^T$. The general case is

far more technical and out of the scope of this paper. We also assume that the point $\theta_\star$ and the r.v. $\boldsymbol{Y}$ satisfy

Assumption 6:

a) $\theta_\star \in \mathcal{L}$.
b) The mean field $h : \mathbb{R}^d \to \mathbb{R}^d$ given by (12) is twice continuously differentiable in a neighborhood of $\theta_\star$.
c) $\nabla h(\theta_\star)$ is a Hurwitz matrix, i.e., the largest real part of its eigenvalues is $-L$ for some $L > 0$.

Assumption 7:

a) There exist $\delta > 0$ and $\tau > 0$ such that $\sup_{|\boldsymbol{\theta}-\mathbf{1}\otimes\theta_\star|\le\delta} \int |\langle \boldsymbol{y}\rangle|^{2+\tau} \mu_{\boldsymbol{\theta}}(d\boldsymbol{y}) < \infty$.
b) The functions $\boldsymbol{\theta} \mapsto \int \langle \boldsymbol{y}\rangle\langle \boldsymbol{y}\rangle^T \mu_{\boldsymbol{\theta}}(d\boldsymbol{y})$ and $\boldsymbol{\theta} \mapsto \int \langle \boldsymbol{y}\rangle \mu_{\boldsymbol{\theta}}(d\boldsymbol{y})$ are continuous in a neighborhood of $\mathbf{1}\otimes\theta_\star$.

We finally strengthen the assumptions on the step-size sequence $(\gamma_n)_{n\ge 0}$. In the sequel, notations $x_n = o(y_n)$ and $x_n \sim y_n$ stand for $x_n/y_n \to 0$ and $x_n/y_n \to 1$, respectively.

Assumption 8:

a) $(\gamma_n)_n$ is a positive deterministic sequence such that either $\log(\gamma_k/\gamma_{k+1}) = o(\gamma_k)$, or $\log(\gamma_k/\gamma_{k+1}) \sim \gamma_k/\gamma_\star$ for some $\gamma_\star > 1/(2L)$.
b) $\sum_n \gamma_n = \infty$ and $\sum_n \gamma_n^2 < \infty$.
c) $\lim_n n\gamma_n = +\infty$ and

$$\lim_n \frac{1}{\sqrt{n}} \sum_{k=1}^n \gamma_k^{-1/2} \left| 1 - \frac{\gamma_k}{\gamma_{k+1}} \right| = 0$$

$$\lim_n \frac{1}{\sqrt{n}} \sum_{k=1}^n \gamma_k = 0.$$

The step size $\gamma_n \sim \gamma_\star/n^\xi$ satisfies Assumptions 8a and 8b for any $1/2 < \xi \le 1$ since $\log(\gamma_k/\gamma_{k+1}) \sim \xi/k$. Similarly, if $\gamma_n \sim \gamma_\star/n$, Assumption 8a holds provided that $\gamma_\star > (1/2L)$. Observe that when the sequence $(\gamma_n)_n$ is ultimately nonincreasing, then the condition $\lim_n n\gamma_n = +\infty$ implies $\lim_n \sqrt{n}^{-1} \sum_{k=1}^n \gamma_k^{-1/2} \ |1 - (\gamma_k/\gamma_{k+1})| = 0$ (see e.g., [21, Th. 26, Ch. 4]). Set

$$\Upsilon := \int \langle \boldsymbol{y}\rangle\langle \boldsymbol{y}\rangle^T \ \mu_{\mathbf{1}\otimes\theta_\star}(d\boldsymbol{y}) - \left(\int \langle \boldsymbol{y}\rangle \ \mu_{\mathbf{1}\otimes\theta_\star}(d\boldsymbol{y}) \right) \left(\int \langle \boldsymbol{y}\rangle \ \mu_{\mathbf{1}\otimes\theta_\star}(d\boldsymbol{y}) \right)^T.$$

Theorem 5: Let us consider Assumptions 1, 4, 6, 7, 8a, and 8b. Assume in addition that $\mathbf{1}^T W_n = \mathbf{1}^T$ w.p.1. Then, under the conditional probability $\mathbb{P}(\cdot | \lim_k \boldsymbol{\theta}_k = \mathbf{1}\otimes\theta_\star)$, the sequence of r.v. $(\gamma_n^{-1/2} (\boldsymbol{\theta}_n - \mathbf{1}\otimes\theta_\star))_{n\ge 0}$ converges in distribution to $\mathbf{1}\otimes Z$ where Z is a centered Gaussian distribution with covariance matrix Σ solution of the Lyapunov equation

$$\nabla h(\theta_\star)\Sigma + \Sigma \nabla h(\theta_\star)^T = -\Upsilon$$

if $\log(\gamma_k/\gamma_{k+1}) = o(\gamma_k)$ and

$$(I + 2\gamma_\star \nabla h(\theta_\star))\,\Sigma + \Sigma\left(I + 2\gamma_\star \nabla h(\theta_\star)^T\right) = -\Upsilon$$

if $\log(\gamma_k/\gamma_{k+1}) \sim \gamma_k/\gamma_\star$.

The proof of Theorem 5 is postponed to Appendix E. The asymptotic variance can be compared to the asymptotic variance in a centralized algorithm: formally, such an algorithm is obtained by setting $W_n = \mathbf{1}\mathbf{1}^T/N \otimes I_d$. Interestingly, the distributed algorithm under study has the same asymptotic variance as its centralized analogue.

Theorem 5 shows that when $\gamma_n \sim \gamma_\star/n^\alpha$ for some $\alpha \in (1/2, 1]$, then the rate in the CLT is $\mathcal{O}(1/n^{\alpha/2})$. Therefore, the maximal rate of convergence is achieved with $\gamma_n \sim \gamma_\star/n$ and in this case, the rate is $\mathcal{O}(1/\sqrt{n})$. Unfortunately, the use of such a rate necessitates to choose $\gamma_\star$ as a function of $\nabla h(\theta_\star)$ (through the upper bound L, see Assumption 8a, and in practice $\nabla h(\theta_\star)$ is unknown. We will show in Theorem 6 that the optimal rate $O(1/\sqrt{n})$ can be reached by applying the averaged procedure (5) with $\gamma_n \sim \gamma_\star/n^\alpha$ whatever $\alpha \in (1/2, 1)$.

A second question is the scaling of the observations in the local step. Observe that during each local step of the algorithm (see (1)), each agent can use a common invertible matrix gain Γ and update the temporary iterate $\tilde{\theta}_{n,i}$ as

$$\tilde{\theta}_{n,i} = \theta_{n-1,i} + \gamma_n \Gamma Y_{n,i}. \tag{19}$$

It is readily seen that the new mean field $\tilde{h} : \theta \mapsto \int \langle (\Gamma \otimes I_N)\boldsymbol{y}\rangle \mu_{\mathbf{1}\otimes\theta}(d\boldsymbol{y})$ is equal to Γh and Assumptions 3 and 4 remain valid with $(\boldsymbol{Y}, h, V)$ replaced by $((\Gamma\otimes I_N)\boldsymbol{Y}, \Gamma h, \Gamma^{-1}V)$. Therefore, introducing a gain matrix Γ does not change the limiting points of the algorithm (4) [and thus (5)] but changes the asymptotic variance. In the case of the optimal rate in Theorem 5 (i.e., the case $\gamma_n \sim \gamma_\star/n$ for some $\gamma_\star > 1/(2L)$), it can be proved following the same lines as in [23] (see also [1, Proposition 4, Ch. 3, Part I]), that the *optimal* choice of the gain matrix is $\Gamma_\star = -\gamma_\star^{-1}\nabla h(\theta_\star)^{-1}$. By optimal, we mean that, when weighting the observations by $\Gamma_\star$ as in (19), the asymptotic covariance matrix $\Sigma_\star$ obtained through Theorem 5 is smaller than the limiting covariance Σ_Γ associated with any other gain matrix Γ, i.e., $\Sigma_\Gamma - \Sigma_\star$ is nonnegative. Moreover, $\Sigma_\star$ is equal to

$$\gamma_\star^{-1} \ \nabla h(\theta_\star)^{-1} \Upsilon \nabla h(\theta_\star)^{-T}.$$

Otherwise stated, $(\sqrt{n} \ \ (\langle \boldsymbol{\theta}_n\rangle - \theta_\star))_{n\ge 0}$ converges to a centered Gaussian vector with covariance matrix $\nabla h(\theta_\star)^{-1}\Upsilon\nabla h(\theta_\star)^{-T}$.

In practice, $\nabla h(\theta_\star)$ is unknown and such a choice of gain matrix cannot be plugged in the algorithm (4). Fortunately, Theorem 6 shows that this optimal variance can be reached by averaging the sequence $(\bar{\boldsymbol{\theta}}_n)_n$.

Note that these two major features of *averaging algorithms* for stochastic approximation (optimal convergence rate and optimal limiting covariance matrix) has been pointed out by [3] (see also [28]) in case of centralized algorithms.

Theorem 6: Let $(\gamma_n)_n$ be a deterministic positive sequence such that $\log(\gamma_k/\gamma_{k+1}) = o(\gamma_k)$. Let us consider Assumptions 1, 4, 6, 7, and 8b–8c. Assume in addition that $\mathbf{1}^T W_n = \mathbf{1}^T$ w.p.1. Then, under the conditional probability $\mathbb{P}(\cdot | \lim_k \boldsymbol{\theta}_k = \mathbf{1}\otimes\theta_\star)$, the sequence of r.v. $(\sqrt{n}\ (\bar{\boldsymbol{\theta}}_n - \mathbf{1}\otimes\theta_\star))_{n\ge 0}$ converges in distribution to $\mathbf{1}\otimes\bar{Z}$ where $\bar{Z}$ is a centered Gaussian distribution with covariance matrix

$$\nabla h(\theta_\star)^{-1} \ \Upsilon \nabla h(\theta_\star)^{-T}.$$

The proof of Theorem 6 is postponed to Appendix F.

V. Application Framework

A. Distributed Estimation

To illustrate the results, we describe in this section a distributed parameter estimation algorithm which converges to a limit point of the centralized ML estimator. Assume that node i receives at time n the $\mathbb{R}^{m_i}$-valued component $X_{n,i}$ of the i.i.d. random process $\boldsymbol{X}_n = (X_{n,1}^T, \ldots X_{n,N}^T)^T \in \mathbb{R}^{\sum m_i}$, where $\boldsymbol{X}_1$ has the unknown density $f_*(x)$ with respect to the Lebesgue measure. The system designer considers that the density of $\boldsymbol{X}_1$ belongs to a family $\{f(\theta, \boldsymbol{x})\}_{\theta\in\mathbb{R}^d}$. When

$f(\theta, \boldsymbol{x})$ satisfies some regularity and smoothness conditions, the limit points of the sequences $\hat{\theta}_n$ that maximize the log-likelihood function $L_n(\theta) = \sum_{k=1}^{n} \log f(\theta, \boldsymbol{X}_k)$ are minimizers of the Kullback–Leibler divergence $D(f_* \,\|\, f(\theta, \cdot))$ [6]. Our aim is to design a distributed and iterative algorithm that exhibits the same asymptotic behavior in the case where $f(\theta, \boldsymbol{x})$ is of the form $f(\theta, \boldsymbol{x}) = \prod_{i=1}^{N} f_i(\theta, x_i)$ where $\boldsymbol{x} = (x_1^T, \ldots, x_N^T)^T$ is partitioned similarly to $\boldsymbol{X}_1$. To that purpose, Algorithm (4) is implemented with the increments $Y_{n+1,i} = \nabla_\theta \log f_i(\theta_{n,i}, X_{n+1,i})$ where ∇_θ is the gradient with respect to θ. In some sense, $\log f_i(\theta_{n,i}, X_{n+1,i})$ is a local log-likelihood function that is updated by node i at time $n+1$ by a gradient approach. Writing $\boldsymbol{\theta} = (\theta_1^T, \ldots, \theta_N^T)^T$, the distribution $\mu_{\boldsymbol{\theta}}$ introduced in Section II-A is defined by the identity

$$\int g(\boldsymbol{y}) \mu_{\boldsymbol{\theta}}(d\boldsymbol{y}) = \int g\big((\nabla_\theta \log f_1(\theta_1, x_1)^T, \ldots \\ \ldots, \nabla_\theta \log f_N(\theta_N, x_N)^T)^T\big)\ f_*(\boldsymbol{x})\, d\boldsymbol{x}$$

for every measurable function $g : \mathbb{R}^{Nd} \to \mathbb{R}_+$. The associated mean field given by (12) will be

$$h(\theta) = \frac{1}{N} \int \nabla_\theta \log f(\theta, \boldsymbol{x})\, f_*(\boldsymbol{x})\, d\boldsymbol{x}.$$

Since $h(\theta) = -N^{-1} \nabla_\theta D(f_* \,\|\, f(\theta, \cdot))$ (assuming ∇_θ and $\int$ can be interchanged), our algorithm is of a gradient type with $V(\theta) = D(f_* \,\|\, f(\theta, \cdot))$ as the natural Lyapunov function. Under the assumptions of Theorems 1 or 2, we know that the $\theta_{n,i}, i = 1, \ldots, N$ converge unanimously to $\mathcal{L} = \{\theta : \nabla V(\theta) = 0\}$. Here, we note that under some weak extra assumptions on the "noise" of the algorithm, it is possible to show that unstable points such as local maxima or saddle points of $V(\theta)$ are avoided (see for instance [29]–[31]). Consequently, the first-order behavior of the distributed algorithm is identical to that of the centralized ML algorithm. We now consider the second-order behavior of these algorithms, restricting ourselves to the case where $f_*(\boldsymbol{x}) = \prod_{i=1}^{N} f_i(\theta_\star, x_i)$ for some $\theta_\star \in \mathbb{R}^d$. With some conditions on f_*, it is well known that any consistent sequence $\hat{\theta}_n$ of estimates provided by the centralized ML algorithm satisfies $\sqrt{n}(\hat{\theta}_n - \theta_\star) \xrightarrow{\mathcal{D}} \mathcal{N}(0, F(\theta_\star)^{-1})$ where $\to \mathcal{D}$ stands for the convergence in distribution, $\mathcal{N}(0, \Sigma)$ represents the centered Gaussian distribution with covariance Σ and

$$F(\theta_\star) = \sum_{i=1}^{N} \int \nabla_\theta \log f_i(\theta_\star, x_i)\, \nabla_\theta \log f_i(\theta_\star, x_i)^T\, f_i(\theta_\star, x_i)\, dx_i$$

is the Fisher information matrix of $f(\theta_\star, \cdot)$ [6, Ch. 6]. We now turn to the distributed algorithm and to that end, we apply Theorems 5 and 6. Matrices $\nabla h(\theta_\star)$ and Υ found in the statements of these theorems coincide in our case with $-N^{-1}F(\theta_\star)$ and $N^{-2}F(\theta_\star)$, respectively (same value of Υ for both theorems). Starting with the averaged case, Theorem 6 shows that on the set $\{\lim_n \boldsymbol{\theta}_n = \mathbf{1} \otimes \theta_\star\}$, the averaged sequence $\bar{\boldsymbol{\theta}}_n$ satisfies $\sqrt{n}(\bar{\boldsymbol{\theta}}_n - \mathbf{1} \otimes \theta_\star) \to \mathcal{D}\mathbf{1} \otimes Z$ where $Z \sim \mathcal{N}(0, F(\theta_\star)^{-1})$. This implies that the averaged algorithm is asymptotically efficient, similarly to the centralized ML algorithm. Let us consider the nonaveraged algorithm. In order to make a fair comparison with the centralized ML algorithm, we restrict the use of Theorem 5 to the case where γ_n has the form $\gamma_n = \gamma_\star / n$. In that case, Assumption 8 is verified when $\gamma_\star > N/(2\lambda_{\min}(F(\theta_\star)))$ where $\lambda_{\min}(F(\theta_\star))$ is the smallest eigenvalue of $F(\theta_\star)$. Theorem 5 shows that on the set $\{\lim_n \boldsymbol{\theta}_n = \mathbf{1} \otimes \theta_\star\}$, the sequence of estimates $\boldsymbol{\theta}_n$ satisfies $\sqrt{n}(\boldsymbol{\theta}_n - \mathbf{1} \otimes \theta_\star) \to \mathcal{D}\mathbf{1} \otimes Z$ where $Z \sim \mathcal{N}(0, \Sigma)$, and where Σ is the solution of the matrix equation $\Sigma(2N^{-1}\gamma_\star F(\theta_\star) - I_d) + (2N^{-1}\gamma_\star F(\theta_\star) - I_d)\Sigma = 2\gamma_\star^2 N^{-2} F(\theta_\star)$. Solving this equation, we obtain $\Sigma = \gamma_\star^2 N^{-2} F(\theta_\star)(2\gamma_\star N^{-1} F(\theta_\star) - I_d)^{-1}$. Notice that $\Sigma - F(\theta_\star)^{-1} = F(\theta_\star)^{-1}(2\gamma_\star N^{-1} F(\theta_\star) - I_d)^{-1}(\gamma_\star N^{-1} F(\theta_\star) - I_d)^2 > 0$, which quantifies the departure from asymptotic efficiency of the nonaveraged algorithm.

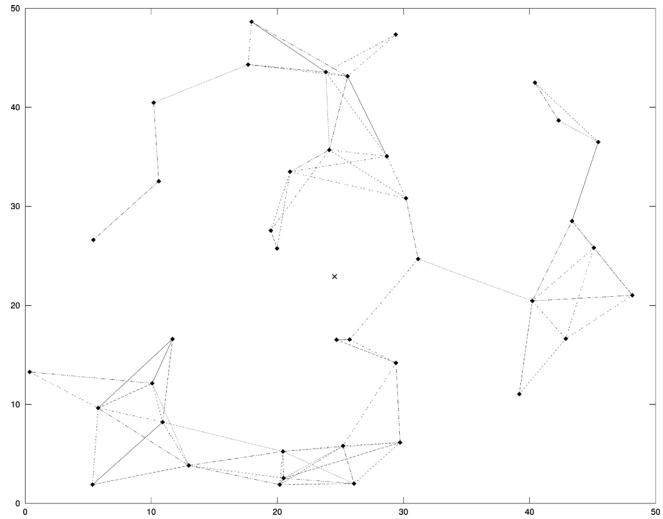

Fig. 1. $N = 40$ sensors with the graph (line segments) and the source (star).

B. Application to Source Localization

The distributed algorithm described above is used here to localize a source by a collection of $N = 40$ sensors. The unknown location of the source in the plane is represented by a parameter $\theta_\star \in \mathbb{R}^2$. The sensors are located in the square $[0, 50] \times [0, 50]$ as shown by Fig. 1, and they receive scalar-valued signals from the source ($m_i = 1$ for all i). It is assumed that the density of $\boldsymbol{X}_1 \in \mathbb{R}^N$ is $f_\star(\boldsymbol{x}) = \prod_{i=1}^{N} f_i(\theta_\star, x_i)$ where $f_i(\theta_\star, \cdot) = \mathcal{N}(1000/|\theta_\star - r_i|^2, 10^{-2})$ where $r_i \in \mathbb{R}^2$ is the location of Node i. The fitted model is $f(\theta, \boldsymbol{x}) = \prod_{i=1}^{N} f_i(\theta, x_i)$ with $f_i(\theta, \cdot) = \mathcal{N}(1000/|\theta - r_i|^2, 10^{-2})$ (see [32] for a similar model). The model for matrices W_n is the pairwise gossip model described in Section II-B. The step sequence γ_n is set to $10^{-3}/n^{0.7}$. Note that in practice, setting adequately the step size in order to find the sought tradeoff between a short transient phase and a good asymptotic accuracy is known to be sensitive to the statistical model of interest. Finally, the initial value $\boldsymbol{\theta}_0 \in \mathbb{R}^{2N}$ is chosen at random under the uniform distribution on the square $[0, 50] \times [0, 50]$.

The convergence of the distributed algorithm to the consensus subspace is illustrated in Fig. 2. Fig. 3 represents the empirical distribution of the normalized estimation error $\gamma_n^{-1/2}(\langle \boldsymbol{\theta}_n \rangle - \theta_\star)$ after $n = 50\,000$ iterations, based on 180 Monte-Carlo runs of the trajectory $\bar{\boldsymbol{\theta}}_n$ initialized in the vicinity of $\theta_\star$. The empirical distribution is coherent with the asymptotic Gaussian distribution given by Theorem 5.

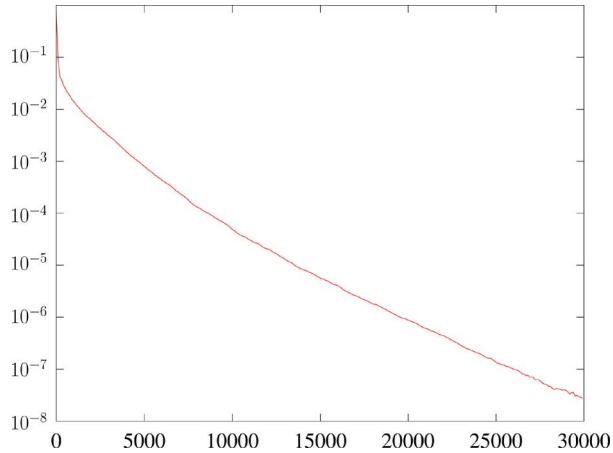

Fig. 2. Square error per node $(1/N)\sum_i |\theta_{n,i} - \theta_\star|^2$ as a function of the number of iterations.

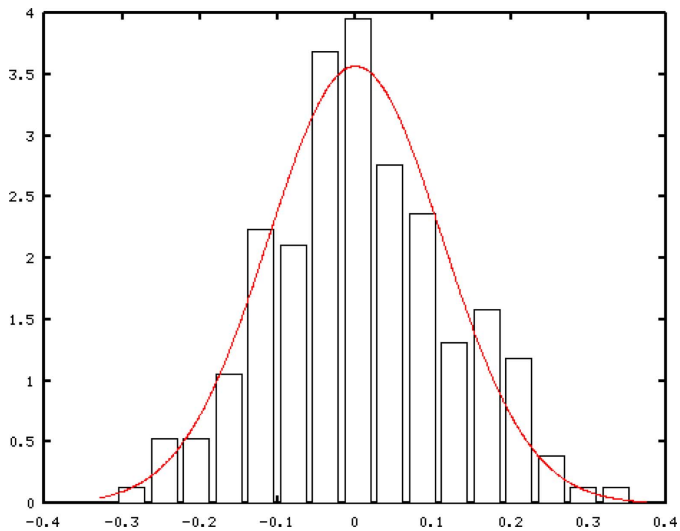

Fig. 3. Empirical distribution of real part of the normalized estimation error $\gamma_n^{-1/2}(\langle \boldsymbol{\theta}_n \rangle - \theta_\star)$ for $n = 50\,000$ (bars) *versus* asymptotic distribution given by Theorem 5 (solid line).

APPENDIX

A. Notations

For a positive deterministic sequence $(a_n)_{n\geq 1}$, the notation $x_n = o(a_n)$ refers to a deterministic $\mathbb{R}^\ell$-valued sequence $(x_n)_{n\geq 1}$ such that $\lim_{n\to\infty} a_n^{-1}|x_n| = 0$. For $p > 0$, we denote the L^p-norm of a random vector X by $\|X\|_p := \mathbb{E}(|X|^p)^{1/p}$. The notation $X_n = o_{L^p}(a_n)$ refers to a $\mathbb{R}^\ell$-valued r.v. $(X_n)_{n\geq 1}$ such that $\lim_{n\to\infty} a_n^{-1}\|X_n\|_p = 0$, while $X_n = \mathcal{O}_{L^p}(a_n)$ refers to a $\mathbb{R}^\ell$-valued r.v. $(X_n)_{n\geq 1}$ such that $\limsup_n a_n^{-1}\|X_n\|_p < \infty$. Finally, $X_n = \mathcal{O}_{w.p.1.}(a_n)$ stands for any $\mathbb{R}^\ell$-valued r.v. $(X_n)_{n\geq 1}$ such that $\limsup_n a_n^{-1}|X_n|$ is finite almost surely.

B. Proof of Theorems 1 and 2

We give the proof of Theorem 2; the proof of Theorem 1 is on the same lines and details are omitted. We first prove the almost sure convergence to zero of $(\boldsymbol{\theta}_{\perp,n})_{n\geq 1}$. The assumption $\mathbb{P}\{\limsup_n |\boldsymbol{\theta}_n| < \infty\} = 1$ implies $\mathbb{P}\left\{\bigcup_{M\in\mathbb{Z}_+}\{\sup_n |\boldsymbol{\theta}_n| \leq M\}\right\} = 1$ and we only have to prove that for any $M > 0$, with probability 1, $\lim_n \boldsymbol{\theta}_{\perp,n}\mathbf{1}_{\sup_n |\boldsymbol{\theta}_n|\leq M} = 0$. To that goal, we write for any $\delta > 0$, $m \geq 1$,

$$\begin{aligned}
&\mathbb{P}\Big\{\sup_{n\geq m} |\boldsymbol{\theta}_{\perp,n}|\mathbf{1}_{\sup_n |\boldsymbol{\theta}_n|\leq M} \geq \delta\Big\} \\
&\quad\leq \frac{1}{\delta^2}\mathbb{E}\left(\sup_{n\geq m} |\boldsymbol{\theta}_{\perp,n}|^2\mathbf{1}_{\sup_n |\boldsymbol{\theta}_n|\leq M}\right) \\
&\quad\leq \frac{1}{\delta^2}\sum_{n\geq m} n^{-2\alpha}\sup_n \mathbb{E}\left(n^{2\alpha}|\boldsymbol{\theta}_{\perp,n}|^2\mathbf{1}_{\sup_{k\leq n-1} |\boldsymbol{\theta}_k|\leq M}\right).
\end{aligned}$$

Lemma 1 and Assumption 5 imply that $(\boldsymbol{\theta}_{\perp,n})_{n\geq 1}$ converges to zero w.p.1. on the set $\{\sup_n |\boldsymbol{\theta}_n| \leq M\}$.

Lemma 1: Let us consider Assumptions 1a, 1b, 2, 4a, and 5. Then, for any $M > 0$,

$$\sup_n n^{2\alpha}\mathbb{E}\left(|\boldsymbol{\theta}_{\perp,n}|^2\,\mathbf{1}_{\sup_{k\leq n-1} |\boldsymbol{\theta}_k|\leq M}\right) < \infty.$$

Proof: Fix $M > 0$. Recalling that $(A\otimes B)(C\otimes D) = (AC)\otimes(BD)$, let $\mathcal{W}_n = (W_n^T\otimes I_d)J_\perp(W_n\otimes I_d) = (W_n^T(I - N^{-1}\mathbf{1}\mathbf{1}^T)W_n)\otimes I_d$. Since $\boldsymbol{\theta}_{\perp,n} = J_\perp(W_n\otimes I_d)(\boldsymbol{\theta}_{\perp,n-1} + \gamma_n\boldsymbol{Y}_n)$, we have by Assumptions 1a and 1b

$$\begin{aligned}
&\mathbb{E}\left[|\boldsymbol{\theta}_{\perp,n}|^2|\mathcal{F}_{n-1}\right] \\
&\quad= \mathbb{E}\left[(\boldsymbol{\theta}_{\perp,n-1} + \gamma_n J_\perp\boldsymbol{Y}_n)^T\mathcal{W}_n(\boldsymbol{\theta}_{\perp,n-1} + \gamma_n\boldsymbol{Y}_n)\,|\mathcal{F}_{n-1}\right] \\
&\quad\leq \rho_n\mathbb{E}\left[|\boldsymbol{\theta}_{\perp,n-1} + \gamma_n\boldsymbol{Y}_n|^2\,|\mathcal{F}_{n-1}\right] \\
&\quad\leq \rho_n\Bigg(|\boldsymbol{\theta}_{\perp,n-1}|^2 + \gamma_n^2\int |\boldsymbol{y}|^2\mu_{\boldsymbol{\theta}_{n-1}}(d\boldsymbol{y}) \\
&\qquad + 2\gamma_n|\boldsymbol{\theta}_{\perp,n-1}|\Big(\int |\boldsymbol{y}|^2\mu_{\boldsymbol{\theta}_{n-1}}(d\boldsymbol{y})\Big)^{1/2}\Bigg).
\end{aligned}$$

By Assumption 4a,

$$\sup_n \int |\boldsymbol{y}|^2\mu_{\boldsymbol{\theta}_{n-1}}(d\boldsymbol{y})\mathbf{1}_{\sup_{k\leq n} |\boldsymbol{\theta}_k|\leq M} < \infty.$$

This implies that there exists a constant $C > 0$ such that

$$\mathbb{E}\left[|\boldsymbol{\theta}_{\perp,n}|^2|\mathcal{F}_{n-1}\right] \leq \rho_n|\boldsymbol{\theta}_{\perp,n-1}|^2 + \gamma_n^2 C + 2\gamma_n|\boldsymbol{\theta}_{\perp,n-1}|\sqrt{C}.$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\left[|\boldsymbol{\theta}_{\perp,n}|^2\mathbf{1}_{\sup_{k\leq n-1} |\boldsymbol{\theta}_k|\leq M}\right] \\
&\quad\leq \rho_n\mathbb{E}\left[|\boldsymbol{\theta}_{\perp,n-1}|^2\mathbf{1}_{\sup_{k\leq n-2} |\boldsymbol{\theta}_k|\leq M}\right] + \gamma_n^2 C \\
&\quad + 2\gamma_n\left(C\,\mathbb{E}\left[|\boldsymbol{\theta}_{\perp,n-1}|^2\mathbf{1}_{\sup_{k\leq n-2} |\boldsymbol{\theta}_k|\leq M}\right]\right)^{1/2}.
\end{aligned}$$

The proof now follows the same lines as in the proof of [33, Lemma 1, Eq. (17)] [see also Lemma 3, (22)]. ■

Remark 2: When Assumption 5 is replaced with Assumption 1c and the condition $\lim_n \gamma_n/\gamma_{n-1} = 1$, then for any $\bar{\rho} \in (\rho, 1)$ there exists a constant C such that

$$\begin{aligned}
&\mathbb{E}\left[\gamma_n^{-2}|\boldsymbol{\theta}_{\perp,n}|^2\mathbf{1}_{\sup_{k\leq n-1} |\boldsymbol{\theta}_k|\leq M}\right] \leq \\
&\qquad\qquad \bar{\rho}\mathbb{E}\left[\gamma_{n-1}^{-2}|\boldsymbol{\theta}_{\perp,n-1}|^2\mathbf{1}_{\sup_{k\leq n-2} |\boldsymbol{\theta}_k|\leq M}\right] + C.
\end{aligned}$$

Therefore, Lemma 1 gets into

$$\sup_n \gamma_n^{-2}\mathbb{E}\left(|\boldsymbol{\theta}_{\perp,n}|^2\,\mathbf{1}_{\sup_{k\leq n-1} |\boldsymbol{\theta}_k|\leq M}\right) < \infty;$$

(see also Theorem 4 for a proof of this bound).

Now, the study of the whole vector $\boldsymbol{\theta}_n$ is reduced to the analysis of its projection $J\boldsymbol{\theta}_n = \mathbf{1} \otimes \langle \boldsymbol{\theta}_n \rangle$ onto the consensus space. We now focus on the average $\langle \boldsymbol{\theta}_n \rangle$. The convergence of the sequence $(\langle \boldsymbol{\theta}_n \rangle)_{n\geq 1}$ is a direct consequence of Lemma 2 along with [22, Ths. 2.2. and 2.3.].

Lemma 2: Under Assumptions 1a, 1b, 2, 4, 5, and (14) it holds

$$\langle \boldsymbol{\theta}_n \rangle = \langle \boldsymbol{\theta}_{n-1} \rangle + \gamma_n h(\langle \boldsymbol{\theta}_{n-1} \rangle) + \gamma_n \zeta_n$$

with $\sup_n |\sum_{k=1}^n \gamma_k \zeta_k| < \infty$ with probability 1. Then, $\lim_n \mathsf{d}(\langle \boldsymbol{\theta}_n \rangle, \mathcal{L}) = 0$ with probability 1.

Proof: Equations (4) and (8) along with Assumption 1a yield

$$\langle \boldsymbol{\theta}_n \rangle = \langle \boldsymbol{\theta}_{n-1} \rangle + \gamma_n \langle \boldsymbol{Z}_n \rangle, \tag{20}$$

where $\boldsymbol{Z}_n := (W_n \otimes I_d)(\boldsymbol{Y}_n + \gamma_n^{-1} \boldsymbol{\theta}_{\perp,n-1})$, upon noting that under Assumption 1a, $(W_n \otimes I_d)J = J$. We write $\langle \boldsymbol{Z}_n \rangle = h(\langle \boldsymbol{\theta}_{n-1} \rangle) + e_n + \xi_n$ where

$$e_n := \langle (W_n \otimes I_d)(\boldsymbol{Y}_n + \gamma_n^{-1} \boldsymbol{\theta}_{\perp,n-1}) \rangle - \int \langle \boldsymbol{y} \rangle \mu_{\boldsymbol{\theta}_{n-1}}(d\boldsymbol{y})$$

$$\xi_n := \int \langle \boldsymbol{y} \rangle \mu_{\boldsymbol{\theta}_{n-1}}(d\boldsymbol{y}) - \int \langle \boldsymbol{y} \rangle \mu_{\mathbf{1} \otimes \langle \boldsymbol{\theta}_{n-1} \rangle}(d\boldsymbol{y}).$$

By Assumption 4b and the inequality $2ab \leq a^2 + b^2$, for any $M > 0$ there exists a constant C such that

$$\mathbb{E}\left| \mathbf{1}_{\sup_n |\boldsymbol{\theta}_n| \leq M} \sum_{n\geq 1} \gamma_n \xi_n \right| \leq C \left(\sum_{n\geq 1} \gamma_n^2 + \sum_{n\geq 1} \mathbb{E}\left(|\boldsymbol{\theta}_{\perp,n-1}|^2 \mathbf{1}_{\sup_n |\boldsymbol{\theta}_n| \leq M} \right) \right). \tag{21}$$

Therefore, the RHS in (21) is finite under the condition 2 and Lemma 1, thus implying that $\sum_{n\geq 1} \gamma_n \xi_n$ converges w.p.1. on the set $\{\sup_n |\boldsymbol{\theta}_n| \leq M\}$ for any $M > 0$ and therefore w.p.1. since $\mathbb{P}\{\sup_n |\boldsymbol{\theta}_n| < \infty\} = 1$.

Since $\mathbb{E}[e_n | \mathcal{F}_{n-1}] = 0$, the sequence $\left(S_n := \sum_{k=1}^n \gamma_k e_k \mathbf{1}_{\sup_{\ell \leq k-1} |\boldsymbol{\theta}_\ell| \leq M} \right)_{n\geq 1}$ is a martingale. We prove that it converges almost surely by estimating its second-order moment. For any $k \geq 1$, see the equation at the bottom of the page, where we set $P_n := N^{-2} W_n^T \mathbf{1}\mathbf{1}^T W_n \otimes I_d$. Note that P_n is independent of Y_n conditionally to $\mathcal{F}_{n-1}$. Since W_n is a stochastic matrix, its spectral norm is bounded uniformly in n. Therefore, there exists a constant $C > 0$ such that

$$\begin{aligned}\mathbb{E}\left[|S_n|^2\right] &\leq C \sum_{n\geq 1} \gamma_n^2 \mathbb{E}\left[\left| \boldsymbol{Y}_n + \gamma_n^{-1} \boldsymbol{\theta}_{\perp,n-1} \right|^2 \mathbf{1}_{\sup_{\ell \leq n-1} |\boldsymbol{\theta}_\ell| \leq M} \right] \\ &\leq 2C \sum_{n\geq 1} \gamma_n^2 \, \mathbb{E}\left[|\boldsymbol{Y}_n|^2 \mathbf{1}_{\sup_{\ell \leq n-1} |\boldsymbol{\theta}_\ell| \leq M} \right] \\ &\quad + 2C \sum_{n\geq 1} \mathbb{E}\left[|\boldsymbol{\theta}_{\perp,n-1}|^2 \mathbf{1}_{\sup_{\ell \leq n-1} |\boldsymbol{\theta}_\ell| \leq M} \right].\end{aligned}$$

By Assumption 4a,

$$\sup_n \mathbb{E}\left[|\boldsymbol{Y}_n|^2 \mathbf{1}_{\sup_{\ell \leq n-1} |\boldsymbol{\theta}_\ell| \leq M} \right] < \infty.$$

By Lemma 1 and Assumption 2, it follows that $\sup_n \mathbb{E}\left[|S_n|^2\right]$ is finite thus implying that the martingale $(S_n)_{n\geq 1}$ converges almost surely to a r.v. which is finite w.p.1. (see e.g., [34, Corollary 2.2.]).

We now consider the last term $\sum_k \gamma_k e_k \left(1 - \mathbf{1}_{\sup_{\ell \leq k-1} |\boldsymbol{\theta}_\ell| \leq M} \right)$. On the set $\{\sup_n |\boldsymbol{\theta}_n| \leq M\}$, this sum is null. This concludes the proof since $\mathbb{P}\{\sup_n |\boldsymbol{\theta}_n| < \infty\} = 1$. ■

C. Proof of Theorem 3

Our stability result relies on preliminary technical lemmas, Lemmas 3 and 4. Theorem 3 is a consequence of Lemma 5: it is established that $\lim_n \boldsymbol{\theta}_{\perp,n} = 0$ with probability 1, which implies that $\mathbb{P}\{\limsup_n |\boldsymbol{\theta}_{\perp,n}| < \infty\} = 1$. It is also established that $\mathbb{P}\{\limsup_n |\langle \boldsymbol{\theta}_n \rangle| < \infty\} = 1$.

Lemma 3: Let $(\gamma_n)_{n\geq 0}, (\rho_n)_{n\geq 0}$ be respectively a positive and a $[0,1]$-valued sequence such that $\sum_n \gamma_n^2 < \infty$; and u_n, v_n be two real sequences such that for $n \geq n_0$,

$$\begin{aligned} u_n &\leq \rho_n u_{n-1} + \gamma_n M \sqrt{u_{n-1}} (1 + u_{n-1} + v_{n-1})^{1/2} \\ &\quad + \gamma_n^2 M (1 + u_{n-1} + v_{n-1}), \end{aligned} \tag{22}$$

$$\begin{aligned} v_n &\leq v_{n-1} + M u_{n-1} + \gamma_n M \sqrt{u_{n-1}} (1 + u_{n-1} + v_{n-1})^{1/2} \\ &\quad + \gamma_n^2 M (1 + u_{n-1} + v_{n-1}). \end{aligned} \tag{23}$$

Then, i) $\sup_n v_n < \infty$, ii) $\limsup_n \phi_n u_n < \infty$ for any positive sequence $(\phi_n)_{n\geq 0}$ such that

$$\limsup_n \left(\gamma_n \sqrt{\phi_n} + \frac{\phi_{n-1}}{\phi_n} \right) < \infty,$$

$$\liminf_n (\gamma_n \sqrt{\phi_n})^{-1} \left(\frac{\phi_{n-1}}{\phi_n} - \rho_n \right) > 0, \tag{24}$$

$$\sum_n \phi_n^{-1} < \infty. \tag{25}$$

$$\begin{aligned}\mathbb{E}\left[|S_k|^2\right] &\leq \sum_{n\geq 1} \gamma_n^2 \, \mathbb{E}\left[|e_n|^2 \mathbf{1}_{\sup_{\ell \leq n-1} |\boldsymbol{\theta}_\ell| \leq M} \right] \\ &\leq \sum_{n\geq 1} \gamma_n^2 \, \mathbb{E}\left[(\boldsymbol{Y}_n + \gamma_n^{-1} \boldsymbol{\theta}_{\perp,n-1})^T P_n (\boldsymbol{Y}_n + \gamma_n^{-1} \boldsymbol{\theta}_{\perp,n-1}) \mathbf{1}_{\sup_{\ell \leq n-1} |\boldsymbol{\theta}_\ell| \leq M} \right]\end{aligned}$$

Proof:

1) Set $\tilde{\gamma}_n = (1+M)\gamma_n$. Define two sequences $(a_n, b_n)_{n\geq n_0}$ such that $a_{n_0} = b_{n_0} = \max(u_{n_0}, v_{n_0})$ and for each $n \geq n_0+1$:

$$\begin{aligned} a_n &= \rho_n a_{n-1} + \tilde{\gamma}_n \sqrt{a_{n-1}}\,(1+a_{n-1}+b_{n-1})^{1/2} \\ &\quad + \tilde{\gamma}_n^2(1+a_{n-1}+b_{n-1}) \qquad (26) \\ b_n &= b_{n-1} + M a_{n-1} + \tilde{\gamma}_n \sqrt{a_{n-1}}(1+a_{n-1}+b_{n-1})^{1/2} \\ &\quad + \tilde{\gamma}_n^2(1+a_{n-1}+b_{n-1}). \qquad (27) \end{aligned}$$

It is straightforward to show by induction that $u_n \leq a_n$ and $v_n \leq b_n$ for any $n \geq n_0$. In addition, $b_n = b_{n-1} + a_n + (M-\rho_n)a_{n-1}$. Thus, for $n \geq n_0+1$,

$$b_n = a_n + \sum_{k=n_0}^{n-1} (M+1-\rho_{k+1})a_k.$$

Define $A_n := (M+1)\sum_{k=n_0}^{n} a_k$, $n \geq n_0$. The above equality implies that $a_n \leq b_n \leq A_n$. As a consequence, (26) implies

$$\begin{aligned} a_n \leq \rho_n a_{n-1} + \tilde{\gamma}_n \sqrt{a_{n-1}}\,(1+2\,A_{n-1})^{1/2} \\ + \tilde{\gamma}_n^2(1+2A_{n-1}). \qquad (28) \end{aligned}$$

As $(A_n)_{n\geq n_0}$ is a positive increasing sequence, for any $n \geq n_0+1$,

$$\begin{aligned} \frac{a_n}{A_n} \leq \rho_n \frac{a_{n-1}}{A_{n-1}} + \tilde{\gamma}_n \sqrt{\frac{a_{n-1}}{A_{n-1}}} \left(\frac{1}{A_{n_0}}+2\right)^{1/2} \\ + \tilde{\gamma}_n^2 \left(\frac{1}{A_{n_0}}+2\right). \qquad (29) \end{aligned}$$

2) Define $L^2 := 1/A_{n_0} + 2$, and $c_n := \phi_n a_n / A_n$. By (29), for any $n \geq n_0+1$,

$$c_n \leq \rho_n \frac{\phi_n}{\phi_{n-1}} c_{n-1} + L\tilde{\gamma}_n \sqrt{c_{n-1}\phi_n} \sqrt{\frac{\phi_n}{\phi_{n-1}}} + L^2\,\tilde{\gamma}_n^2 \phi_n, \quad (30)$$

and under the assumption (24), there exist $n_1 \geq n_0$ and a constant $\xi > 0$ such that for any $n \geq n_1$,

$$\begin{aligned} \sqrt{\frac{\phi_{n-1}}{\phi_n}} L\xi \left\{1 + \xi L \tilde{\gamma}_n \sqrt{\phi_{n-1}}\right\} \\ \leq \left(\frac{\phi_{n-1}}{\phi_n} - \rho_n\right) \left(\tilde{\gamma}_n \sqrt{\phi_n}\right)^{-1}. \qquad (31) \end{aligned}$$

Define

$$A := \max\left(\frac{1}{\xi}, \frac{1}{\xi^2}, c_{n_1}\right). \qquad (32)$$

We prove by induction on n that $c_n \leq A$ for any $n \geq n_1$. The claim holds true for $n = n_1$ by definition of A. Assume that $c_{n-1} \leq A$ for some $n-1 \geq n_1$. Using (30) and (32), for $n \geq n_1+1$,

$$\frac{c_n}{A} \leq \rho_n \frac{\phi_n}{\phi_{n-1}} + \frac{L}{\sqrt{A}} \tilde{\gamma}_n \sqrt{\phi_n} \sqrt{\frac{\phi_n}{\phi_{n-1}}} + \frac{L^2}{A}\,\tilde{\gamma}_n^2 \phi_n.$$

By (31), the RHS is less than one so that $c_n \leq A$. This proves that $(c_n)_{n\geq n_0}$ is a bounded sequence.

3) We prove that $(A_n)_{n\geq n_0}$ is a bounded sequence. Using the fact that $\sup_{n\geq n_1} \rho_n \leq 1$, $(A_n)_{n\geq n_0}$ is increasing and (28), it holds for $n \geq n_1+1$

$$\begin{aligned} A_n &= A_{n-1} + a_n \\ &\leq A_{n-1} + a_{n-1} + \tilde{\gamma}_n \sqrt{a_{n-1}}\,\sqrt{A_{n-1}} L^{1/2} + \tilde{\gamma}_n^2 L^2 A_{n-1} \\ &\leq \left(1 + c_{n-1}\phi_{n-1}^{-1} + L^{1/2}\tilde{\gamma}_n \phi_{n-1}^{-1/2} \sqrt{c_{n-1}} + \tilde{\gamma}_n^2 L^2\right) A_{n-1}. \end{aligned}$$

Finally, since $\sup_{n\geq n_1} c_n \leq A$ and $(1+t^2) \leq \exp(t^2)$, there exists $C > 0$ s.t. for any $n \geq n_1+1$, $A_n \leq \exp\left(C\{\phi_{n-1}^{-1} + \tilde{\gamma}_n^2\}\right) A_{n-1}$ (note that under (24), $\limsup_n \{\tilde{\gamma}_n/\sqrt{\phi_n}\}\phi_n < \infty$). By assumptions, $\sum_n \{\phi_{n-1}^{-1} + \tilde{\gamma}_n^2\} < \infty$, $(A_n)_{n\geq n_0}$ is therefore bounded.

4) The proof of the lemma is concluded upon noting that $v_n \leq b_n \leq A_n$ and $u_n \leq a_n \leq \tilde{\gamma}_n^2 c_n A_n$. ■

Remark 3: If the sequences $(\gamma_n, \rho_n)_{n\geq 0}$ are such that

$$\begin{aligned} &\limsup_n \left(\frac{\gamma_n}{\gamma_{n-1}} + \frac{1-\rho_{n-1}}{1-\rho_n}\right) < \infty, \\ &\liminf_n \frac{1}{1-\rho_n} \left(\frac{(1-\rho_{n-1})^2}{(1-\rho_n)^2} \frac{\gamma_n^2}{\gamma_{n-1}^2} - \rho_n\right) > 0, \qquad (33) \\ &\sum_n \gamma_n^2 (1-\rho_n)^{-2} < \infty, \qquad (34) \end{aligned}$$

then the conditions (24) and (25) are satisfied with $\phi_n := (1-\rho_n)^2/\gamma_n^2$. Examples of sequences satisfying these conditions are $\rho_n = 1 - a/n^\eta$, $\gamma_n = \gamma_0/n^\xi$ with $0 \leq \eta < 1 \wedge (\xi - 1/2)$.

Lemma 4: Let $V : \mathbb{R}^d \to \mathbb{R}^+$ be a differentiable function such that ∇V is Lipschitz on $\mathbb{R}^d$. There exist constants C, C' such that for any $\theta \in \mathbb{R}^d$, $|\nabla V(\theta)|^2 \leq C V(\theta)$, and for any θ, $\theta' \in \mathbb{R}^d$,

$$V(\theta') \leq V(\theta) + \nabla V(\theta)^T(\theta'-\theta) + C'|\theta'-\theta|^2. \qquad (35)$$

Proof: Given any θ, $\theta' \in \mathbb{R}^d$, we have

$$\begin{aligned} V(\theta') = V(\theta) + \nabla V(\theta)^T(\theta'-\theta) \\ + \int_0^1 \left(\nabla V(\theta + t(\theta'-\theta)) - \nabla V(\theta)\right)^T (\theta'-\theta)\, dt. \end{aligned}$$

This implies (35) since ∇V is Lipschitz. Then, applying (35) with $\theta' = \theta - \mu \nabla V(\theta)$ where $\mu > 0$ and recalling that V is nonnegative, we also have $0 \leq V(\theta) - \mu(1-\mu C')|\nabla V(\theta)|^2$. Choosing μ small enough, we thus get the result. ■

Lemma 5 (Agreement and Stability): Let us consider Assumptions 1a, 1b, 2, 3a, 3b, and 5. Assume in addition ST1 and ST2. Then,

a) $\sum_{n\geq 1} \mathbb{E}\,|\boldsymbol{\theta}_{\perp,n}|^2 < \infty$ and $(\boldsymbol{\theta}_{\perp,n})_{n\geq 1}$ converges to zero w.p.1.

b) $\sup_{n\geq 1} \mathbb{E} V(\langle \boldsymbol{\theta}_n \rangle) < \infty$ and $\sup_n \mathbb{E}\left[|\boldsymbol{Y}_n|^2\right] < \infty$,

where $\langle \boldsymbol{x} \rangle$ and $\boldsymbol{x}_\perp$ are given by (8) and (9).

Proof: Define $u_n := \mathbb{E}\left[|\boldsymbol{\theta}_{\perp,n}|^2\right]$ and $v_n := \mathbb{E}\left[V(\langle \boldsymbol{\theta}_n \rangle)\right]$. We prove that there exists a constant $M > 0$ and an integer n_0 such that for any $n \geq n_0$, inequalities (22) and (23) are satisfied. The proof is then concluded by application of Lemma 3 upon noting that under Assumption 2, the rate $\phi_n = n^{2\alpha}$ satisfies the conditions (24) and (25).

Proof of (22): As $W_n \mathbf{1} = \mathbf{1}$, we have $J_\perp(W_n \otimes I_d) = J_\perp(W_n \otimes I_d) J_\perp$. As a consequence, $\boldsymbol{\theta}_{\perp,n} = J_\perp(W_n \otimes$

$I_d)(\boldsymbol{\theta}_{\perp,n-1}+\gamma_n \boldsymbol{Y}_n)$. We expand the square Euclidean norm of the latter vector (see the equation given at the bottom of page), integrate both sides of the above equation w.r.t. the r.v. W_n; by Assumption 1b

$$\mathbb{E}[|\boldsymbol{\theta}_{\perp,n}|^2 \,|\mathcal{F}_{n-1}, \boldsymbol{Y}_n] \leq \rho_n |\boldsymbol{\theta}_{\perp,n-1}+\gamma_n \boldsymbol{Y}_n|^2.$$

Under Assumption 5, $\lim_n n(1-\rho_n) = +\infty$: then, there exists n_0 such that $\rho_n < 1$ for any $n \geq n_0$. We obtain

$$\mathbb{E}[|\boldsymbol{\theta}_{\perp,n}|^2] \leq \rho_n \mathbb{E}[|\boldsymbol{\theta}_{\perp,n-1}|^2] + 2\gamma_n \mathbb{E}[|\boldsymbol{\theta}_{\perp,n-1}|\,|\boldsymbol{Y}_n|] + \gamma_n^2 \mathbb{E}[|\boldsymbol{Y}_n|^2],$$

for any $n \geq n_0$. From Cauchy–Schwartz inequality, $\mathbb{E}[|\boldsymbol{\theta}_{\perp,n-1}|\,|\boldsymbol{Y}_n|] \leq \sqrt{u_{n-1}}(\mathbb{E}[|\boldsymbol{Y}_n|^2])^{1/2}$. Thus,

$$u_n \leq \rho_n u_{n-1} + 2\gamma_n \sqrt{u_{n-1}}(\mathbb{E}[|\boldsymbol{Y}_n|^2])^{1/2} + \gamma_n^2 \mathbb{E}[|\boldsymbol{Y}_n|^2].$$

By Assumption ST2, we have the following estimate $\mathbb{E}[|\boldsymbol{Y}_n|^2] \leq C_1(1+v_{n-1}+u_{n-1})$. This completes the proof of (22), for any constant M larger than $1+C_1$.

Proof of (23): Lemma 4 is applied with $\theta \leftarrow \langle \boldsymbol{\theta}_n \rangle$ and $\theta' \leftarrow \langle \boldsymbol{\theta}_{n-1} \rangle$. We have to evaluate the difference $\langle \boldsymbol{\theta}_n \rangle - \langle \boldsymbol{\theta}_{n-1} \rangle$. By (4),

$$\langle \boldsymbol{\theta}_n \rangle = \left(\frac{\mathbf{1}^T W_n}{N} \otimes I_d\right)(\boldsymbol{\theta}_{n-1} + \gamma_n \boldsymbol{Y}_n).$$

Therefore,

$$\begin{aligned}\langle \boldsymbol{\theta}_n \rangle - \langle \boldsymbol{\theta}_{n-1} \rangle &= \left(\frac{\mathbf{1}^T W_n - \mathbf{1}^T}{N} \otimes I_d\right)\boldsymbol{\theta}_{n-1} + \left(\frac{\mathbf{1}^T W_n}{N} \otimes I_d\right)\gamma_n \boldsymbol{Y}_n \\ &= \left(\frac{\mathbf{1}^T W_n - \mathbf{1}^T}{N} \otimes I_d\right)\boldsymbol{\theta}_{\perp,n-1} + \left(\frac{\mathbf{1}^T W_n}{N} \otimes I_d\right)\gamma_n \boldsymbol{Y}_n, \qquad (36)\end{aligned}$$

where the second equality is due to the fact that W_n is row-stochastic. Under Assumption 1a, $\mathbb{E}(W_n)$ is doubly stochastic. Thus, using the Assumption 1b

$$\mathbb{E}[\langle \boldsymbol{\theta}_n \rangle - \langle \boldsymbol{\theta}_{n-1} \rangle | \mathcal{F}_{n-1}] = \gamma_n \int \langle \boldsymbol{y} \rangle \, \mu_{\boldsymbol{\theta}_{n-1}}(d\boldsymbol{y}). \qquad (37)$$

Plugging (37) into (35), there exists C' such that for any n,

$$\begin{aligned}\mathbb{E}[V(\langle \boldsymbol{\theta}_n \rangle)|\mathcal{F}_{n-1}] \leq\; & V(\langle \boldsymbol{\theta}_{n-1} \rangle) \\ & + \gamma_n \nabla V(\langle \boldsymbol{\theta}_{n-1} \rangle)^T \int \langle \boldsymbol{y} \rangle \, \mu_{\boldsymbol{\theta}_{n-1}}(d\boldsymbol{y}) \\ & + C' \mathbb{E}[|\langle \boldsymbol{\theta}_n \rangle - \langle \boldsymbol{\theta}_{n-1} \rangle|^2 | \mathcal{F}_{n-1}].\end{aligned}$$

By the Condition 3b, the quantity $-\nabla V(\langle \boldsymbol{\theta}_{n-1} \rangle)^T h(\langle \boldsymbol{\theta}_{n-1} \rangle)$ is positive; therefore,

$$\begin{aligned}\mathbb{E}[V(\langle \boldsymbol{\theta}_n \rangle)|\mathcal{F}_{n-1}] \leq\; & V(\langle \boldsymbol{\theta}_{n-1} \rangle) \\ & + \gamma_n \nabla V(\langle \boldsymbol{\theta}_{n-1} \rangle)^T \left(\int \langle \boldsymbol{y} \rangle \, \mu_{\boldsymbol{\theta}_{n-1}}(d\boldsymbol{y}) - h(\langle \boldsymbol{\theta}_{n-1} \rangle) \right) \\ & + C' \mathbb{E}[|\langle \boldsymbol{\theta}_n \rangle - \langle \boldsymbol{\theta}_{n-1} \rangle|^2 | \mathcal{F}_{n-1}].\end{aligned}$$

Using successively the Conditions ST2 and Lemma 4, we have the estimate

$$\begin{aligned}&\nabla V(\langle \boldsymbol{\theta}_{n-1} \rangle)^T \left(\int \langle \boldsymbol{y} \rangle \, \mu_{\boldsymbol{\theta}_{n-1}}(d\boldsymbol{y}) - h(\langle \boldsymbol{\theta}_{n-1} \rangle) \right) \\ &\quad \leq |\nabla V(\langle \boldsymbol{\theta}_{n-1} \rangle)| \, C_2 |\boldsymbol{\theta}_{\perp,n-1}| \\ &\quad \leq \sqrt{C} C_2 \sqrt{V(\langle \boldsymbol{\theta}_{n-1} \rangle)} \, |\boldsymbol{\theta}_{\perp,n-1}|.\end{aligned}$$

Using Cauchy–Schwartz inequality, the expectation of the above quantity is no larger than $\sqrt{C} C_2 \sqrt{u_{n-1} v_{n-1}}$. We obtain

$$v_n \leq v_{n-1} + \gamma_n \sqrt{C} C_2 \sqrt{u_{n-1}(1+u_{n-1}+v_{n-1})} + C' \mathbb{E}[|\langle \boldsymbol{\theta}_n \rangle - \langle \boldsymbol{\theta}_{n-1} \rangle|^2], \qquad (38)$$

where we used the fact that $u_{n-1} \geq 0$. We now need to find an estimate for $\mathbb{E}[|\langle \boldsymbol{\theta}_n \rangle - \langle \boldsymbol{\theta}_{n-1} \rangle|^2]$. Using Minkowski's inequality on (36),

$$\begin{aligned}&\mathbb{E}[|\langle \boldsymbol{\theta}_n \rangle - \langle \boldsymbol{\theta}_{n-1} \rangle|^2]^{1/2} \\ &\quad \leq \mathbb{E}\left[\left|\left(\frac{\mathbf{1}^T W_n - \mathbf{1}^T}{N} \otimes I_d\right)\boldsymbol{\theta}_{\perp,n-1}\right|^2\right]^{1/2} \\ &\qquad + \mathbb{E}\left[\left|\left(\frac{\mathbf{1}^T W_n}{N} \otimes I_d\right)\gamma_n \boldsymbol{Y}_n\right|^2\right]^{1/2}. \qquad (39)\end{aligned}$$

Focus on the first term of the RHS of the above inequality. Remark that

$$\mathbb{E}[(W_n^T \mathbf{1} - \mathbf{1})(\mathbf{1}^T W_n - \mathbf{1}^T)|\mathcal{F}_{n-1}] = \mathbb{E}[W_n^T \mathbf{1}\mathbf{1}^T W_n] - \mathbf{1}\mathbf{1}^T,$$

where we used the Assumption 1b along with the fact that $\mathbb{E}(W_n)$ is doubly stochastic (see the condition 1a)). Upon noting that the entries of W_n are in $[0,1]$ (as a consequence of Assumption 1a, the spectral norm of $\mathbb{E}[W_n^T \mathbf{1}\mathbf{1}^T W_n] - \mathbf{1}\mathbf{1}^T$ is bounded. Thus, there exists a constant C' such that

$$\mathbb{E}\left[\left|\left(\frac{\mathbf{1}^T W_n - \mathbf{1}^T}{N} \otimes I_d\right)\boldsymbol{\theta}_{\perp,n-1}\right|^2\right] \leq C' u_{n-1}.$$

$$|\boldsymbol{\theta}_{\perp,n}|^2 = (\boldsymbol{\theta}_{\perp,n-1} + \gamma_n \boldsymbol{Y}_n)^T (\{W_n^T (I_N - \mathbf{1}\mathbf{1}^T/N) W_n\} \otimes I_d)(\boldsymbol{\theta}_{\perp,n-1} + \gamma_n \boldsymbol{Y}_n)$$

By similar arguments, there exists a constant C'' such that

$$\mathbb{E}\left[\left|\left(\frac{\mathbf{1}^T W_n}{N}\otimes I_d\right)\gamma_n \boldsymbol{Y}_n\right|^2\right] \le C''\gamma_n^2\,\mathbb{E}|\boldsymbol{Y}_n|^2 \le C_2 C''\gamma_n^2\,(1+u_{n-1}+v_{n-1})$$

where we used Assumption ST2. Putting this together with (39),

$$\begin{aligned}&\mathbb{E}[|\langle\boldsymbol{\theta}_n\rangle - \langle\boldsymbol{\theta}_{n-1}\rangle|^2]\\ &\quad\le (\sqrt{C'}\sqrt{u_{n-1}} + \gamma_n\sqrt{C_2C''}\,\sqrt{1+u_{n-1}+v_{n-1}})^2\\ &\quad\le C(u_{n-1}+\gamma_n^2\,(1+u_{n-1}+v_{n-1})\\ &\quad+\gamma_n\sqrt{u_{n-1}(1+u_{n-1}+v_{n-1})})\end{aligned}$$

where $C>0$ is some constant chosen large enough. Plugging the above inequality into (38),

$$\begin{aligned}v_n &\le v_{n-1} + (C'C)u_{n-1}\\ &\quad+(\sqrt{C}C_2+C'C)\gamma_n\sqrt{u_{n-1}(1+u_{n-1}+v_{n-1})}\\ &\quad+C'C\gamma_n^2\,(1+u_{n-1}+v_{n-1}).\end{aligned}$$

This proves that (23) holds for any M chosen large enough.

Proof of $\sup_n \mathbb{E}\left[|\boldsymbol{Y}_n|^2\right] < \infty$*:* By Assumptions 1b and ST2:

$$\begin{aligned}\mathbb{E}\left[|\boldsymbol{Y}_n|^2\right] &= \mathbb{E}\left[\mathbb{E}_{\boldsymbol{\theta}_{n-1}}\left[|\boldsymbol{Y}|^2\right]\right]\\ &\le C_2\left(1+\mathbb{E}\left[V(\langle\boldsymbol{\theta}_{n-1}\rangle)\right]+\mathbb{E}\left[|\boldsymbol{\theta}_{\perp,n-1}|^2\right]\right).\end{aligned}\quad(40)$$

The proof follows since $\sup_n \mathbb{E}\left[V(\langle\boldsymbol{\theta}_n\rangle)\right] < \infty$ and $\mathbb{E}\left[|\boldsymbol{\theta}_{\perp,n}|^2\right] \le \sum_n \mathbb{E}\left[|\boldsymbol{\theta}_{\perp,n}|^2\right] < \infty$. ■

Lemma 6: Let us consider Assumptions 1a, 1b, 2, 3a–3e, and 5. Assume in addition ST1 and ST2. Then, $\mathbb{P}\left\{\limsup_n |\langle\boldsymbol{\theta}_n\rangle| < \infty\right\} = 1$.

Proof: The sequence $(\langle\boldsymbol{\theta}_n\rangle)_{n\ge1}$ satisfies the (10). The proof is an application of [22, Th. 2.2.]: in order to apply this theorem, we only have to prove that with probability 1 (i) the sequence $(\langle\boldsymbol{\theta}_n\rangle)_{n\ge1}$ is infinitely often in a level set $\{V\le M\}$ i.e., $\mathbb{P}\left\{\liminf_n V(\langle\boldsymbol{\theta}_n\rangle) < \infty\right\} = 1$ and (ii)

$$\sum_n \gamma_n\left((W_n\otimes I_d)(\boldsymbol{Y}_n+\gamma_n^{-1}\boldsymbol{\theta}_{\perp,n-1}) - h(\langle\boldsymbol{\theta}_{n-1}\rangle)\right) < \infty.$$

For the recurrence property, we have

$$\begin{aligned}\mathbb{E}\left(\liminf_n V(\langle\boldsymbol{\theta}_n\rangle)\right) &\le \liminf_n \mathbb{E}\left(V(\langle\boldsymbol{\theta}_n\rangle)\right)\\ &\le \sup_n \mathbb{E}\left(V(\langle\boldsymbol{\theta}_n\rangle)\right).\end{aligned}$$

By Lemma 5, the RHS is finite thus showing that $\mathbb{P}\left\{\liminf_n V(\langle\boldsymbol{\theta}_n\rangle) < \infty\right\} = 1$. For the second property, we write $\langle(W_n\otimes I_d)(\boldsymbol{Y}_n+\gamma_n^{-1}\boldsymbol{\theta}_{\perp,n-1})\rangle - h(\langle\boldsymbol{\theta}_{n-1}\rangle) = e_n+\xi_{n-1}$ where

$$e_n := \langle(W_n\otimes I_d)(\boldsymbol{Y}_n+\gamma_n^{-1}\boldsymbol{\theta}_{\perp,n-1})\rangle - \int\langle\boldsymbol{y}\rangle\mu_{\boldsymbol{\theta}_{n-1}}(d\boldsymbol{y})$$

$$\xi_{n-1} := \int\langle\boldsymbol{y}\rangle\mu_{\boldsymbol{\theta}_{n-1}}(d\boldsymbol{y}) - \int\langle\boldsymbol{y}\rangle\mu_{\mathbf{1}\otimes\langle\boldsymbol{\theta}_{n-1}\rangle}(d\boldsymbol{y}).$$

By Assumption ST2 and the inequality $2ab \le a^2+b^2$, there exists a constant C such that

$$\mathbb{E}\left|\sum_{n\ge1}\gamma_n\xi_{n-1}\right| \le C\left(\sum_{n\ge1}\gamma_n^2 + \sum_{n\ge1}\mathbb{E}\,|\boldsymbol{\theta}_{\perp,n-1}|^2\right).\quad(41)$$

Therefore, the RHS in (41) is finite under the condition 2 and Lemma 5, thus implying that $\sum_{n\ge1}\gamma_n\xi_n$ converges w.p.1. Since $\mathbb{E}\left[e_n\,|\mathcal{F}_{n-1}\right]=0$, the sequence $(S_n := \sum_{k=1}^n \gamma_k e_k)_{n\ge1}$ is a martingale. We prove that it converges almost surely by estimating its second-order moment. For any $k\ge1$,

$$\begin{aligned}\mathbb{E}\left[|S_k|^2\right] &\le \sum_{n\ge1}\gamma_n^2\,\mathbb{E}\left[|e_n|^2\right]\\ &\le \sum_{n\ge1}\gamma_n^2\,\mathbb{E}\left[(\boldsymbol{Y}_n+\gamma_n^{-1}\boldsymbol{\theta}_{\perp,n-1})^T P_n(\boldsymbol{Y}_n+\gamma_n^{-1}\boldsymbol{\theta}_{\perp,n-1})\right]\end{aligned}$$

where we set $P_n := N^{-2}W_n^T\mathbf{1}\mathbf{1}^T W_n\otimes I_d$. Note that P_n is independent of Y_n conditionally to $\mathcal{F}_{n-1}$. Since W_n is a stochastic matrix, its spectral norm is bounded uniformly in n. Therefore, there exists a constant $C>0$ such that

$$\begin{aligned}\mathbb{E}\left[|S_n|^2\right] &\le C\sum_{n\ge1}\gamma_n^2\,\mathbb{E}\left[\left|\boldsymbol{Y}_n+\gamma_n^{-1}\boldsymbol{\theta}_{\perp,n-1}\right|^2\right]\\ &\le 2C\sum_{n\ge1}\gamma_n^2\,\mathbb{E}\left[|\boldsymbol{Y}_n|^2\right] + 2C\sum_{n\ge1}\mathbb{E}\left[|\boldsymbol{\theta}_{\perp,n-1}|^2\right].\end{aligned}$$

By Lemma 5 and Assumption 2 it follows that $\sup_n \mathbb{E}\left[|S_n|^2\right]$ is finite thus implying that the martingale $(S_n)_{n\ge1}$ converges almost surely to a r.v. which is finite w.p.1. (see e.g., [34, Corollary 2.2.]). This concludes the proof. ■

D. Proof of Theorem 4

Set $V_n := (I_N - \mathbf{1}\mathbf{1}^T/N)W_n$ and for any $1\le k\le n$,

$$\Phi_{n,k} := (V_n\otimes I_d)(V_{n-1}\otimes I_d)\cdots(V_k\otimes I_d).\quad(42)$$

Note that by Assumptions 1b–1c,

$$\begin{aligned}\|\Phi_{n,k}X\|_2^2 &= \mathbb{E}[X^T\Phi_{n-1,k}^T(V_n^T V_n\otimes I_d)\Phi_{n-1,k}X]\\ &= \mathbb{E}[X^T\Phi_{n-1,k}^T\mathbb{E}(V_n^T V_n\otimes I_d)\Phi_{n-1,k}X]\\ &\le \rho\,\mathbb{E}[X^T\Phi_{n-1,k}^T\Phi_{n-1,k}X] = \rho\|\Phi_{n-1,k}X\|_2^2.\end{aligned}\quad(43)$$

From (4) and since $J_\perp(W_n\otimes I_d) = J_\perp(W_n\otimes I_d)J_\perp = (V_n\otimes I_d)J_\perp$ by Assumption 1a, it holds for any $n\ge1$, $\boldsymbol{\theta}_{\perp,n} = (V_n\otimes I_d)(\boldsymbol{\theta}_{\perp,n-1}+\gamma_n\boldsymbol{Y}_{\perp,n})$. By induction,

$$\boldsymbol{\theta}_{\perp,n} = \sum_{k=1}^n \gamma_k\Phi_{n,k}\boldsymbol{Y}_{\perp,k} + \Phi_{n,1}\boldsymbol{\theta}_{\perp,0}\quad(44)$$

where $\Phi_{n,k}$ is defined by (42). By (43) and Assumption 1c, the second term in the RHS of (44) is a $\mathcal{O}_{L^2}(\rho^{n/2})$. We now

consider the first term in the RHS of (44). Using Minkowski's inequality and (43)

$$\begin{aligned}\|\sum_{k=1}^{n}\gamma_k\Phi_{n,k}\boldsymbol{Y}_{\perp,k}\mathbf{1}_{\sup_{\ell\le n-1}|\boldsymbol{\theta}_\ell|\le M}\|_2 &\le \sum_{k=1}^{n}\gamma_k\|\Phi_{n,k}\boldsymbol{Y}_{\perp,k}\mathbf{1}_{\sup_{\ell\le n-1}|\boldsymbol{\theta}_\ell|\le M}\|_2\\ &\le \sum_{k=1}^{n}\gamma_k\sqrt{\rho}^{n-k+1}\|\boldsymbol{Y}_{\perp,k}\mathbf{1}_{\sup_{\ell\le k-1}|\boldsymbol{\theta}_\ell|\le M}\|_2.\end{aligned}$$

By [35, Result 178, p. 38], the RHS is upper bounded by $\limsup_{n\to\infty}\|\boldsymbol{Y}_{\perp,n}\mathbf{1}_{|\boldsymbol{\theta}_{n-1}|\le M}\|_2\rho(1-\sqrt{\rho})^{-1}$. Under Assumption 4a, this upper bound is finite (the proof follows the same lines as in the proof of Lemma 2 and is omitted). This concludes the proof.

E. Proof of Theorem 5

Assumption 2 implies that $\lim_n \rho^{n/2}\gamma_n^{-2}=0$. Upon noting that

$$\mathbb{P}\left\{\bigcup_M\{\sup_n|\boldsymbol{\theta}_n|\le M\}|\lim_q\boldsymbol{\theta}_q=\mathbf{1}\otimes\theta_\star\right\}=1,$$

Theorem 4 implies that the sequence of r.v. $(\gamma_n^{-1/2}\boldsymbol{\theta}_{\perp,n})_n$ converges in probability to zero under the conditional probability $\mathbb{P}\{\cdot|\lim_q\boldsymbol{\theta}_q=\mathbf{1}\otimes\theta_\star\}$. Since $\boldsymbol{\theta}_n=\mathbf{1}\otimes\langle\boldsymbol{\theta}_n\rangle+\boldsymbol{\theta}_{\perp,n}$, it remains to prove that the sequence of r.v. $(\gamma_n^{-1/2}(\langle\boldsymbol{\theta}_n\rangle-\theta_\star))_{n\ge0}$ converges in distribution to Z under the conditional distribution given the event $\{\lim_q\theta_q=\mathbf{1}\otimes\theta_\star\}$. To that goal, we write

$$\langle\boldsymbol{\theta}_n\rangle=\langle\boldsymbol{\theta}_{n-1}\rangle+\gamma_n h(\langle\boldsymbol{\theta}_{n-1}\rangle)+\gamma_n e_n+\gamma_n\xi_n$$

where $\xi_n:=\int\langle\boldsymbol{y}\rangle\mu_{\boldsymbol{\theta}_{n-1}}(d\boldsymbol{y})-\int\langle\boldsymbol{y}\rangle\mu_{\mathbf{1}\otimes\langle\boldsymbol{\theta}_{n-1}\rangle}(d\boldsymbol{y})$ and

$$\begin{aligned}e_n&:=\langle(W_n\otimes I_d)(\boldsymbol{Y}_n+\gamma_n^{-1}\boldsymbol{\theta}_{\perp,n-1})\rangle\\&\quad-\int\langle\boldsymbol{y}\rangle\mu_{\boldsymbol{\theta}_{n-1}}(d\boldsymbol{y})=\langle\boldsymbol{Y}_n\rangle-\int\langle\boldsymbol{y}\rangle\mu_{\boldsymbol{\theta}_{n-1}}(d\boldsymbol{y}),\end{aligned}$$

since $\mathbf{1}^TW_n=\mathbf{1}^T$. We then check the conditions C1 to C4 of [23, Th. 1] (see also [24, Th. 1]). Under the Assumptions 6 and 8a, the conditions C1 and C4 of [23, Th. 1] are satisfied. We now prove C2b: there exists a constant C such that

$$\begin{aligned}&\mathbb{E}\left[|e_{n+1}|^{2+\tau}\mathbf{1}_{|\boldsymbol{\theta}_n-\mathbf{1}\otimes\theta_\star|\le\delta}\right]\\&\le C\,\mathbb{E}\left[|\int\langle\boldsymbol{y}\rangle\mu_{\boldsymbol{\theta}_n}(d\boldsymbol{y})|^{2+\tau}\mathbf{1}_{|\boldsymbol{\theta}_n-\mathbf{1}\otimes\theta_\star|\le\delta}\right]\\&\quad+C\,\mathbb{E}\left[|\langle\boldsymbol{Y}_{n+1}\rangle|^{2+\tau}\mathbf{1}_{|\boldsymbol{\theta}_n-\mathbf{1}\otimes\theta_\star|\le\delta}\right]\\&\le 2C\sup_{|\boldsymbol{\theta}-\mathbf{1}\otimes\theta_\star|\le\delta}\int|\langle\boldsymbol{y}\rangle|^{2+\tau}\mu_{\boldsymbol{\theta}}(d\boldsymbol{y})\end{aligned}$$

and the RHS is finite under Assumption 7. For C2c, we have

$$\begin{aligned}&\mathbb{E}\left[e_{n+1}e_{n+1}^T|\mathcal{F}_n\right]\\&=\left\{\int\langle\boldsymbol{y}\rangle\langle\boldsymbol{y}\rangle^T\mu_{\boldsymbol{\theta}_n}(d\boldsymbol{y})-\left(\int\langle\boldsymbol{y}\rangle\mu_{\boldsymbol{\theta}_n}(d\boldsymbol{y})\right)\left(\int\langle\boldsymbol{y}\rangle\mu_{\boldsymbol{\theta}_n}(d\boldsymbol{y})\right)^T\right\}.\end{aligned}$$

By Assumption 7, this term converges w.p.1 to Υ on the set $\{\lim_k\boldsymbol{\theta}_k=\mathbf{1}\otimes\theta_\star\}$. This concludes the proof of C2.

We now consider the condition C3 of [23] with $r_n=\xi_n+e_n\mathbf{1}_{|\boldsymbol{\theta}_{n-1}-\mathbf{1}\otimes\theta_\star|>\delta}$: we prove that for any $M>0$, $\gamma_n^{-1/2}r_n\mathbf{1}_{\sup_k|\boldsymbol{\theta}_k|\le M}\mathbf{1}_{\lim_k\boldsymbol{\theta}_k=\mathbf{1}\otimes\theta_\star}=\mathcal{O}_{w.p.1}o_{L^1}(1)$. By Assumption 4b, there exists a constant C such that

$$\begin{aligned}&\gamma_n^{-1/2}\mathbb{E}\left[|\xi_n|\mathbf{1}_{\lim_k\boldsymbol{\theta}_k=\mathbf{1}\otimes\theta_\star}\mathbf{1}_{\sup_k|\boldsymbol{\theta}_k|\le M}\right]\\&\qquad\le C\,\left(\gamma_n^{-1}\mathbb{E}\left[|\boldsymbol{\theta}_{\perp,n}|^2\mathbf{1}_{\sup_k|\boldsymbol{\theta}_k|\le M}\right]\right)^{1/2}\end{aligned}$$

and the RHS tends to zero as $n\to\infty$ by Theorem 4. On the set $\{\lim_n\boldsymbol{\theta}_n=\mathbf{1}\otimes\theta_\star\}$, the r.v. $e_n\mathbf{1}_{|\boldsymbol{\theta}_{n-1}-\mathbf{1}\otimes\theta_\star|>\delta}$ is null for all large n. This concludes the proof of the condition C3 of [23], and the proof of Theorem 5.

F. Proof of Theorem 6

We preface the proof by a preliminary result, established by [23, Th. 2] (see also [21] for a similar result obtained under stronger assumptions).

Theorem 7: Let $(\gamma_n)_n$ be a deterministic positive sequence such that $\log(\gamma_k/\gamma_{k+1})=o(\gamma_k)$ and satisfying Assumptions 8b and 8c. Consider the random sequence $(u_n)_n$ given by

$$u_{n+1}=u_n+\gamma_{n+1}h(u_n)+\gamma_{n+1}e_{n+1}+\gamma_{n+1}\xi_{n+1},\quad u_0\in\mathbb{R}^d,$$

where

AVER1. $u_\star$ is a zero of the mean field: $h(u_\star)=0$. The mean field $h:\mathbb{R}^d\to\mathbb{R}^d$ is twice continuously differentiable (in a neighborhood of $u_\star$) and $\nabla h(u_\star)$ is a Hurwitz matrix.

AVER2.

(i) $(e_n)_{n\ge1}$ is a $\mathcal{F}_n$-adapted martingale-increment sequence.

(ii) For any $M>0$, there exist $\tau>0$ s.t. $\sup_k\mathbb{E}\left[|e_k|^{2+\tau}\mathbf{1}_{\sup_{\ell\le k-1}|u_\ell-u_\star|\le M}\right]<\infty$.

(iii) There exists a positive definite (random) matrix $U_\star$ such that on the set $\{\lim_q u_q=u_\star\}$, $\lim_k\mathbb{E}\left[e_ke_k^T|\mathcal{F}_{k-1}\right]=U_\star$ almost surely.

AVER3. $(\xi_n)_{n\ge1}$ is a $\mathcal{F}_n$-adapted sequence s.t.

(i) $\gamma_n^{-1/2}|\xi_n|\mathbf{1}_{\lim_q u_q=u_\star}\mathbf{1}_{\sup_n|u_n|\le M}=\mathcal{O}_{w.p.1}(1)\mathcal{O}_{L^2}(1)$ for any $M>0$.

(ii) $n^{-1/2}\sum_{k=0}^{n}\xi_{k+1}\mathbf{1}_{\lim_q u_q=u_\star}$ converges to zero in probability.

Then, for any $t\in\mathbb{R}^d$,

$$\begin{aligned}&\lim_n\mathbb{E}\left[\mathbf{1}_{\lim_q u_q=\theta_\star}\,\exp\left(i\sqrt{n}\,t^T\left(\frac{1}{n}\sum_{k=1}^{n}u_k-u_\star\right)\right)\right]\\&=\mathbb{E}\left[\mathbf{1}_{\lim_q u_q=u_\star}\,\exp\left(-\frac{1}{2}t^T\nabla h(u_\star)^{-1}\,U_\star\,\nabla h(u_\star)^{-T}t\right)\right].\end{aligned}$$

Proof of Theorem 6: By Theorem 4 and Assumption 8c, $\sqrt{N}^{-1}\sum_{n=1}^{N}\boldsymbol{\theta}_{\perp,n}\mathbf{1}_{\sup_\ell|\boldsymbol{\theta}_\ell|\le M}$ converges in L^2 to zero for any $M>0$. Since $\boldsymbol{\theta}_n=\boldsymbol{\theta}_{\perp,n}+\mathbf{1}\otimes\langle\boldsymbol{\theta}_n\rangle$, we now prove a CLT for the averaged sequence $N^{-1}\sum_{n=1}^{N}\langle\boldsymbol{\theta}_n\rangle$. To that goal, we check the Assumptions AVER 1 to AVER 3 of Theorem 7 with $u_n=\langle\boldsymbol{\theta}_n\rangle$; e_n,ξ_n defined as in the Proof of Theorem 5. AVER 1 and AVER 2 can be proved along the same lines as in the proof of Theorem

5; details are omitted. Finally, by Assumption 4b and Theorem 4, $\mathbb{E}\left[|\xi_n|^2 \mathbf{1}_{\lim_k \boldsymbol{\theta}_k = \mathbf{1}\otimes\theta_\star} \mathbf{1}_{\sup_{\ell\le n-1}|\boldsymbol{\theta}_\ell|\le M}\right] = \mathcal{O}(\gamma_n^2)$; and

$$\ell^{-1/2}\sum_{n=1}^{\ell}\mathbb{E}\left[|\xi_n|\mathbf{1}_{\lim_k \boldsymbol{\theta}_k = \mathbf{1}\otimes\theta_\star} \mathbf{1}_{\sup_{\ell\le n-1}|\boldsymbol{\theta}_\ell|\le M}\right] \le C\,\ell^{-1/2}\sum_{n=1}^{\ell}\gamma_n.$$

The RHS tends to zero under Assumption 8c thus showing AVER 3.

REFERENCES

[1] A. Benveniste, M. Metivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*. New York, NY, USA: Springer-Verlag, 1987.

[2] H. J. Kushner and G. G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*. New York, NY, USA: Springer-Verlag, 2003.

[3] B. Polyak, "New stochastic approximation type procedures," *Autom. Remote Control*, vol. 51, pp. 98–107, 1990.

[4] P. Forero, A. Cano, and G. Giannakis, "Consensus-based distributed support vector machines," *J. Mach. Learning Res.*, vol. 11, pp. 1663–1707, 2010.

[5] P. Bianchi and J. Jakubowicz, "On the convergence of a multi-agent projected stochastic gradient algorithm for non convex optimization," *IEEE Trans. Autom. Control*, vol. 58, no. 2, pp. 391–405, Feb. 2013, arXiv:1107.2526v1.

[6] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, ser. Springer Texts in Statistics, second ed. New York, NY, USA: Springer-Verlag, 1998.

[7] V. D. Blondel, J. M. Hendrickx, A. Olshevsky, and J. N. Tsitsiklis, "Convergence in multiagent coordination, consensus, and flocking," in *Proc. 44th IEEE Conf. Decision Control Eur. Control Conf.*, Dec. 2005, pp. 2996–3000.

[8] C. Lopes and A. H. Sayed, "Distributed processing over adaptive networks," in *Proc. Adapt. Sens. Array Process. Workshop*, Jun. 2006, pp. 1–5.

[9] S. Kar and J. M. F. Moura, "Distributed consensus algorithms in sensor networks: Quantized data and random link failures," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1383–1400, Mar. 2010.

[10] F. Cattivelli and A. H. Sayed, "Diffusion LMS strategies for distributed estimation," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1035–1048, Mar. 2010.

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

[12] A. Nedic, "Asynchronous broadcast-based convex optimization over a network," *IEEE Trans. Autom. Control*, vol. 56, no. 6, pp. 1337–1351, Jun. 2011.

[13] S. S. Stanković, M. S. Stanković, and D. M. Stipanović, "Decentralized parameter estimation by consensus based stochastic approximation," *IEEE Trans. Autom. Control*, vol. 56, no. 3, pp. 531–543, Mar. 2011.

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

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

[16] H. J. Kushner and G. Yin, "Asymptotic properties of distributed and communicating stochastic approximation algorithms," *SIAM J. Control Optim.*, vol. 25, pp. 1266–1290, 1987.

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

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

[19] S. Ram, A. Nedic, and V. Veeravalli, "Distributed stochastic subgradient projection algorithms for convex optimization," *J. Optim. Theory Appl.*, vol. 147, pp. 516–545, 2010, 10.1007/s10957-010-9737-7.

[20] P. Bianchi, G. Fort, W. Hachem, and J. Jakubowicz, "Performance of a distributed Robbins-Monro algorithm for sensor networks," presented at the 19th European Signal Processing Conf., Barcelona, Spain, 2011.

[21] B. Delyon, Stochastic approximation with decreasing gain: convergence and asymptotic theory. , Unpublished Lecture Notes, 2000 [Online]. Available: http://perso.univ-rennes1.fr/bernard.delyon/as_cours.ps

[22] C. Andrieu, E. Moulines, and P. Priouret, "Stability of stochastic approximation under verifiable conditions," *SIAM J. Control Optim.*, vol. 44, no. 1, pp. 283–312, 2005.

[23] G. Fort, "A central limit theorem for a stochastic approximation algorithm and its Polyak-averaged version," 2012 [Online]. Available: http://perso.telecom-paristech.fr/~gfort/Preprints/CLTforSA.pdf

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

[25] F. Bénézit, "Distributed average consensus for wireless sensor networks," Ph.D. dissertation, École Polytechnique Fédérale de Lausanne, , 2009.

[26] T. C. Aysal, M. E. Yildiz, A. D. Sarwate, and A. Scaglione, "Broadcast gossip algorithms for consensus," *IEEE Trans. Signal Process.*, vol. 57, no. 7, pp. 2748–2761, Jul. 2009.

[27] H. Chen, L. Guo, and A. Gao, "Convergence and robustness of the Robbins-Monro algorithm truncated at randomly varying bounds," *Stochast. Process. Appl.*, vol. 27, pp. 217–231, 1988.

[28] B. T. Polyak and A. B. Juditsky, "Acceleration of stochastic approximation by averaging," *SIAM J. Control Optim.*, vol. 30, pp. 838–855, 1992.

[29] O. Brandière and M. Duflo, "Les algorithmes stochastiques contournent-ils les pièges?," *Ann. Inst. H. Poincaré Probab. Statist.*, vol. 32, no. 3, pp. 395–427, 1996.

[30] M. Benaim, "Dynamics of stochastic approximation algorithms," in *Séminaire de Probabilités*, ser. Lecture Notes in Math. Berlin, Germany: Springer-Verlag, 1999, vol. 1709, pp. 1–68.

[31] H.-T. Fang and H.-F. Chen, "Stability and instability of limit points for stochastic approximation algorithms," *IEEE Trans. Autom. Control*, vol. 45, no. 3, pp. 413–420, Mar. 2000.

[32] M. Rabbat and R. Nowak, "Distributed optimization in sensor networks," in *Proc. 3rd Int. Symp. Inf. Process. Sens. Netw.*, 2004, pp. 20–27.

[33] P. Bianchi and J. Jakubowicz, "Distributed stochastic approximation for constrained and unconstrained optimization," in *Proc. 5th Int. ICST Conf. Performance Evaluation Methodologies Tools*, 2011, pp. 227–233.

[34] P. Hall and C. C. Heyde, *Martingale Limit Theory and its Application*. New York, NY, USA: Academic, 1980.

[35] G. Pólya and G. Szegő, *Problems and Theorems in Analysis*, ser. Integral Calculus. Theory of Functions. New York, NY, USA: Springer-Verlag (Classics in Mathematics), 1998.

Pascal Bianchi (M'12) was born in 1977 in Nancy, France. He received the M.Sc. degree of the University of Paris XI and Supélec in 2000 and the Ph.D. degree of the University of Marne-la-Vallée in 2003. From 2003 to 2009, he was an Associate Professor at the Telecommunication Department of Supélec. In 2009, he joined the Statistics and Applications group at LTCI-Telecom ParisTech. His current research interests are in the area of distributed algorithms for multi-agent networks. They include distributed optimization, stochastic approximation, decentralized detection, quantization.

Gersende Fort was born in France, in 1974. She received the Engineering degree in Telecommunications from Ecole Nationale Supérieure des Télécommunications (ENST), Paris, France, in 1997; the Master's degree from Université Paris IV, France, in 1997; the PhD degree in Applied Mathematics from Université Paris IV in 2001; and the Habilitationà Diriger les Recherches from Université Dauphine-Paris IX in 2010.

She joined the Centre national de la Recherche Scientifique (CNRS) in 2001 and she is now a senior research scientist at Laboratoire Traitement et Communication de l'Information (LTCI). Her research interests are on Bayesian inverse problems and Monte Carlo methods.

She served as an associate editor for *Bernoulli Journal* since 2013.

Walid Hachem (M'04) was born in Bhamdoun, Lebanon, in 1967. He received the Engineering degree in telecommunications from St Joseph University (ESIB), Beirut, Lebanon, in 1989, the Master's degree from Telecom ParisTech, France, in 1990, the Ph.D. degree in signal processing from Université Paris Est Marne-La-Vallée in 2000 and he Habilitationà Diriger des Recherches from Université Paris-Sud in 2006.

After working in the telecommunications industry for ten years, he joined the academia in 2001 as a faculty member at Supélec, France. In 2006, he joined the CNRS (Centre national de la Recherche Scientifique), where he is now a research director at Telecom ParisTech. His research themes consist mainly in the random matrix theory and its applications in signal processing, and in the decentralized estimation and optimization algorithms.

He served as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING between 2007 and 2010.