Performance of a Distributed Stochastic Approximation Algorithm

Pascal Bianchi, *Member, IEEE*, Gersende Fort, 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.

I. Introduction

Stochastic approximation has been a very active research area for the last sixty 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 a function 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

The authors are with LTCI - CNRS/TELECOM ParisTech, 46 rue Barrault, 75634 Paris Cedex 13, France (e-mail: {name}@telecom-paristech.fr). This work is partially supported by the French National Research Agency under the program ANR-07 ROBO 002.

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, ...). 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 j's and computes the weighted average:

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

where the $w_n(i,j)$'s are scalar non-negative 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. These matrices are called row-stochastic, since they have non negative elements and satisfy $W_n \mathbf{1} = \mathbf{1}$ where $\mathbf{1}$ is the $N \times 1$ vector whose components are all equal to one.

This paper analyzes the convergence of this algorithm under some mild assumptions. In particular, due to the matrices W_n, the estimates will eventually reach the *consensus* in the sense that the differences $\theta_{n,i} - \theta_{n,j}$ between the estimates of any two nodes i and j almost surely converge to zero as $n \to \infty$. Asymptotic fluctuations of the estimates will also be studied through Central Limit Theorems.

There is a rich literature on distributed estimation and optimization algorithms, see [3],[4], [5], [6], [7], [8] as a non exhaustive list. Among the first gossip algorithms are those considered in the treatise [9] and in [10]. The case where the gossip matrices are random and the observations are noiseless is considered in [11]. The authors of [7] solve a constrained optimization by also using noiseless estimates. The contributions [6] and [8] consider the framework of linear regression

models. In [12], stochastic gradient algorithms are considered in the case the matrices $(W_n)_n$ are doubly stochastic gossip *i.e.* $W_n\mathbf{1} = W_n^T\mathbf{1} = \mathbf{1}$. This contribution assumes in addition that the gradients are bounded and considers rather stringent assumptions on the conditional variances of the observation noises.

The contributions of this paper are summarized as follows:

- The distributed stochastic approximation algorithm introduced above is studied under very general assumptions. In particular, the algorithm is not required to be of gradient type. Stability and convergence are established with the help of a Lyapunov function. It is shown that the sequences of estimates at all nodes converge unanimously to an equilibrium set of the noiseless recursion seen as a dynamical system.
- The random gossip matrices W_n are assumed to be row stochastic and, column stochastic in the mean, *i.e.*, $W_n\mathbf{1} = \mathbf{1}$ and $\mathbf{1}^T\mathbb{E}[W_n] = \mathbf{1}^T$. 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 $\mathbf{1}^TW_n = \mathbf{1}^T$ which is assumed in many contributions (see *e.g.* [13], [7], [12], [14]) 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 [15] (see also Section II-B below) are only column stochastic in the mean.
- 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.
- Finally, 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 analogue.
- We also consider a CLT on the sequences averaged over time as introduced in [16]. We show that averaging always improves the rate of convergence and the asymptotic variance.

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. Section VI is devoted to the proofs. An application relative to distributed estimation is described in Section V, along with some numerical simulations. The appendix contains some technical details.

II. The 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:

$$\boldsymbol{\theta}_n = (W_n \otimes I_d)\left(\boldsymbol{\theta}_{n-1} + \gamma_n \boldsymbol{Y}_n\right) \ , \tag{2}$$

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 [16], we also consider the *averaged* sequence $(\bar{\boldsymbol{\theta}}_n)_{n\geq 1}$ given by

$$\bar{\theta}_{n,i} = \frac{1}{n}\sum_{k=1}^{n} \theta_{k,i} \tag{3}$$

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. Similarly, we note $\bar{\boldsymbol{\theta}}_n := (\bar{\theta}_{n,1}^T, \ldots, \bar{\theta}_{n,N}^T)^T$. 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 $[0,1]$. For any $\boldsymbol{\theta} \in \mathbb{R}^{dN}$, we denote by $\mathbb{E}_{\boldsymbol{\theta}}$ the expectation with respect to (w.r.t.) the distribution $\mu_{\boldsymbol{\theta}}$.

We consider the case when the random variables (r.v.) $(\boldsymbol{Y}_n, W_n)_{n\geq 1}$ are defined on 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 non-negative elements such that:

- W_n is row stochastic: $W_n \mathbf{1} = \mathbf{1}$,
- $\mathbb{E}(W_n)$ is column stochastic: $\mathbf{1}^T \mathbb{E}(W_n) = \mathbf{1}^T$,

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

$$\mathbb{E}[f(W_{n+1}) g(\boldsymbol{Y}_{n+1}) | \mathcal{F}_n] = \mathbb{E}[f(W_{n+1})] \, \mathbb{E}_{\boldsymbol{\theta}_n}[g(\boldsymbol{Y})] \ . \tag{4}$$

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 [11], random gossip is assumed in this paper. Assumption **1a**) has been commented in the introduction. 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 *(i)* the r.v. W_n and Y_n are independent conditionally to the past, *(ii)* the r.v. $(W_n)_{n\geq 1}$ are independent and *(iii)* the conditional distribution of $\boldsymbol{Y}_{n+1}$ given the past is $\mu_{\boldsymbol{\theta}_n}$.

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 $\lim_n \gamma_n/\gamma_{n+1} = 1$, $\sum_n \gamma_n = \infty$ and $\sum_n \gamma_n^2 < \infty$.

B. Illustration: Some Examples of Gossip Schemes

We describe two standard gossip schemes so called *pairwise* and *broadcast* schemes. The reader can refer to [17] for a more complete picture and for more general gossip strategies. The network of agents is represented as a non-directed 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 [11] on average consensus (see also [18]).

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 [11]: 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 [15]. 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 [15]). 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 [15]).

III. Convergence results

In this section, we address the asymptotic behavior when $n \to \infty$ of the algorithm (2) and of its averaged version (3). We prove in Theorem 1 that all agents eventually reach an agreement on the value of their estimate: the limit points of $(\boldsymbol{\theta}_n)_{n\geq 1}$ (resp. $(\bar{\boldsymbol{\theta}}_n)_{n\geq 1}$) given by (2) (resp. (3)) are of the form $\mathbf{1} \otimes \theta_\star$.

A. Notations

Denote by $|x|$ the Euclidean norm of a vector x and by ∇ the gradient operator (on $\mathbb{R}^d$). Let

$$J := (\mathbf{1}\mathbf{1}^T/N) \otimes I_d \ , \qquad J_\perp := I_{dN} - J \ , \tag{5}$$

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

$$\langle\boldsymbol{x}\rangle := \frac{1}{N}(\mathbf{1}^T\otimes I_d)\boldsymbol{x}\ , \tag{6}$$

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

$$\boldsymbol{x}_\perp := J_\perp\boldsymbol{x} \tag{7}$$

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*.

B. Assumptions on the distributions $\mu_{\boldsymbol{\theta}}$

In order to derive the convergence results, assumptions on the probability measures $(\mu_{\boldsymbol{\theta}})_{\boldsymbol{\theta}\in\mathbb{R}^{dN}}$ have to be introduced. Define the function $h:\mathbb{R}^d\to\mathbb{R}^d$ by:

$$h(\theta) := \mathbb{E}_{\mathbf{1}\otimes\theta}\left[\langle\boldsymbol{Y}\rangle\right]\ . \tag{8}$$

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 $\nabla V^T\ h\le 0$.

It is assumed:

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

a) V is differentiable and ∇V is a Lipschitz function.

b) For any $\theta\in\mathbb{R}^d$, $\nabla V(\theta)^T h(\theta)\le 0$, where h is given by (8).

c) There exists a constant C_1, such that for any $\theta\in\mathbb{R}^d$, $|\nabla V(\theta)|^2\le C_1(1+V(\theta))$.

d) For any $M>0$, the level set $\{\theta\in\mathbb{R}^d : V(\theta)\le M\}$ is compact.

e) The set $\mathcal{L} := \{\theta\in\mathbb{R}^d : \nabla V(\theta)^T h(\theta) = 0\}$ is non-empty and bounded.

f) $V(\mathcal{L})$ has an empty interior.

Assumption **3c)** implies that the Lyapunov function V increases at most at a quadratic rate when $|\theta|\to\infty$. Assumption **3f)** is 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, the Sard's theorem implies that $g(\{\nabla g=0\})$ has an empty interior. If g is strictly convex with optimum $\theta_\star$, the function $\theta\mapsto|\theta-\theta_\star|^2$ is also a Lyapunov function. In this case, $\mathcal{L}=\{\theta_\star\}$.

Assumption 4: a) There exists a constant C_2 such that for any $\boldsymbol{\theta} \in \mathbb{R}^{dN}$,

$$\mathbb{E}_{\boldsymbol{\theta}} \left[|\boldsymbol{Y}|^2 \right] \leq C_2 \left(1 + V(\langle \boldsymbol{\theta} \rangle) + |\boldsymbol{\theta}_\perp|^2 \right) \ , \tag{9}$$

$$\left| \mathbb{E}_{\boldsymbol{\theta}} \langle \boldsymbol{Y} \rangle - \mathbb{E}_{\mathbf{1} \otimes \langle \boldsymbol{\theta} \rangle} \langle \boldsymbol{Y} \rangle \right| \leq C_2 |\boldsymbol{\theta}_\perp| \ . \tag{10}$$

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

Condition (9) implies that $|h(\theta)|^2 \leq C_2(1 + V(\theta))$ (set $\boldsymbol{\theta} = \mathbf{1} \otimes \theta$ and use Jensen's inequality). Combined with assumption **3**, this means that $h(\theta)$ is at most linearly increasing when $|\theta| \to \infty$.

C. 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: Under Assumptions **1**, **2**, **3** and **4**, w.p.1,

$$\lim_{n \to \infty} \mathsf{d}(\langle \boldsymbol{\theta}_n \rangle, \mathcal{L}) = 0 \ , \qquad \lim_{n} \boldsymbol{\theta}_{\perp,n} = 0 \ , \tag{11}$$

where $\mathcal{L}$ is given by Assumption **3**. Moreover, w.p.1, $(\langle \boldsymbol{\theta}_n \rangle)_{n \geq 1}$ converges to a connected component of $\mathcal{L}$.

Theorem 1 states that, almost surely, the vector of iterates $\boldsymbol{\theta}_n$ given by (2) converges to the consensus space as $n \to \infty$ so that the network asymptotically achieves consensus.

The assumptions of Theorem 1 imply that w.p.1, the sequence $\{V(\langle \boldsymbol{\theta}_n \rangle)\}_{n \geq 0}$ converges to a (random) point $v_\star \in V(\mathcal{L})$. This can be used to show that $(\langle \boldsymbol{\theta}_n \rangle)_{n \geq 0}$ converges to a connected component of $\{\theta \in \mathcal{L} : V(\theta) = v_\star\}$. In general, this does not imply that $(\langle \boldsymbol{\theta}_n \rangle)_{n \geq 0}$ 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 $(\boldsymbol{\theta}_n)_{n \geq 0}$ converging to $\mathbf{1} \otimes \theta_\star$ for some $\theta_\star \in \mathcal{L}$, the 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 (3) and the original sequence (2) have the same limiting value, if any.

D. Case of a vanishing communication rate

Theorem 1 still holds true when the r.v. $(W_n)_{n \geq 1}$ are not identically distributed. An interesting example is when W_n is the identity matrix with a probability that tends to one 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 (2) and (3) when the spectral norm ρ_n of the matrix 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:

$$\lim_{n\to\infty} n^\alpha \gamma_n = 0\,, \qquad \lim_{n\to\infty} n^{1+\alpha}\gamma_n = +\infty\ , \tag{12}$$

$$\liminf_{n\to\infty} \frac{1-\rho_n}{n^\alpha \gamma_n} > 0\ , \tag{13}$$

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 $\liminf_n p_n/(n^\alpha\gamma_n) = +\infty$: 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 (13) 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

Theorem 2: The statement of Theorem 1 remains valid under Assumptions **1a-b)**, **3**, **4** and **5**.
Theorem 2 is proved in Section VI-C.

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 (5) and (7)). 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 $\boldsymbol{\theta}_{\perp,n} \to 0$ almost surely, Theorem 3 provides an information on the convergence rate: $\boldsymbol{\theta}_{\perp,n}$ tends to zero in L^2 at rate $1/\gamma_n$.

Theorem 3: Under Assumptions **1**, **2**, **3** and **4**,

$$\gamma_n^{-2}\mathbb{E}\left(|\boldsymbol{\theta}_{\perp,n}|^2\right) \leq \frac{\rho\ C}{(1-\sqrt{\rho})^2} + \mathcal{O}\left(\rho^{n/2}\gamma_n^{-2}\right) \tag{14}$$

where ρ is given by Assumption **1c)** and $C := \limsup_{n\to\infty} \mathbb{E}(|\boldsymbol{Y}_{\perp,n}|^2)$ 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. $\mathbf{1}^T W_n = \mathbf{1}^T$. 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 (8) 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 $\sup_{|\boldsymbol{\theta}-\mathbf{1}\otimes\theta_\star|\leq\delta} \mathbb{E}_{\boldsymbol{\theta}}\left[|\langle \boldsymbol{Y}\rangle|^{2+\tau}\right] < \infty$.

b) The function $\boldsymbol{\theta} \mapsto \mathbb{E}_{\boldsymbol{\theta}}\left[\langle \boldsymbol{Y}\rangle\langle \boldsymbol{Y}\rangle^T\right]$ is continuous in a neighborhood of $\mathbf{1}\otimes\theta_\star$.

We finally strengthen the assumptions on the step-size sequence $(\gamma_n)_{n\geq 0}$ and assume that

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

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

The step size $\gamma_n \sim \gamma_\star/n^\xi$ satisfies Assumptions **8a-b)** for any $1/2 < \xi \leq 1$ 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 non-increasing, 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. [19, Theorem 26, Chapter 4]).

Set

$$\Upsilon := \mathbb{E}_{\mathbf{1}\otimes\theta_\star}\left[\langle \boldsymbol{Y}\rangle\langle \boldsymbol{Y}\rangle^T\right] - \mathbb{E}_{\mathbf{1}\otimes\theta_\star}\left[\langle \boldsymbol{Y}\rangle\right]\mathbb{E}_{\mathbf{1}\otimes\theta_\star}\left[\langle \boldsymbol{Y}\rangle\right]^T \ .$$

Theorem 4: Let Assumptions **1**, **3**, **4**, **6, 7, 8a-b)** hold true. 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. $(\gamma_n^{-1/2} \ (\boldsymbol{\theta}_n - \mathbf{1}\otimes\theta_\star))_{n\geq 0}$ converges in distribution to $\mathbf{1}\otimes Z$ where Z is a centered Gaussian distribution with covariance matrix Σ solution of the Lyapunov equation:

$$\begin{cases} \nabla h(\theta_\star)\Sigma + \Sigma\nabla h(\theta_\star)^T = -\Upsilon & \text{if } \log(\gamma_k/\gamma_{k+1}) = o(\gamma_k) \ , \\ \left(I + 2\gamma_\star\nabla h(\theta_\star)\right)\Sigma + \Sigma\left(I + 2\gamma_\star\nabla h(\theta_\star)^T\right) = -\Upsilon & \text{if } \log(\gamma_k/\gamma_{k+1}) \sim \gamma_k/\gamma_\star \ . \end{cases}$$

The proof of Theorem 4 is postponed to Section VI-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 4 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 5 that the optimal rate $O(1/\sqrt{n})$ can be reached by applying the averaged procedure (3) 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

$$\tilde{\theta}_{n,i} = \theta_{n-1,i} + \gamma_n \, \Gamma Y_{n,i} \ . \tag{15}$$

It is readily seen that the new mean field $\tilde{h} : \theta \mapsto \mathbb{E}_{\mathbf{1}\otimes\theta}[\langle (\Gamma\otimes I_N)\boldsymbol{Y}\rangle]$ 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 (2) (and thus (3)) but changes the asymptotic variance. In the case of the optimal rate in Theorem 4 (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 [20]

(see also [1, Proposition 4, Chapter 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 (15), the asymptotic covariance matrix $\Sigma_\star$ obtained through Theorem 4 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, $(\sqrt{n}\ (\langle\boldsymbol{\theta}_n\rangle - \theta_\star))_{n\geq 0}$ 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 (2). Fortunately, Theorem 5 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 [16] (see also [21]) in case of centralized algorithms.

Theorem 5: Let $(\gamma_n)_n$ be a deterministic positive sequence such that $\log(\gamma_k/\gamma_{k+1}) = o(\gamma_k)$. Let Assumptions **1**, **3**, **4**, **6**, **7**, **8b-c)** hold true. 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. $(\sqrt{n}\ (\bar{\boldsymbol{\theta}}_n - \mathbf{1}\otimes\theta_\star))_{n\geq 0}$ converges 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 is postponed to Section VI-F.

V. An 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 Maximum Likelihood (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 $D(f_* \,\|\, f(\theta,\cdot))$ [22]. 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 (2) is implemented with the increments $Y_{n+1,i} = \nabla_\theta \log f_i\left(\theta_{n,i}, X_{n+1,i}\right)$ 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

$$\mathbb{E}_{\boldsymbol{\theta}}[g(\boldsymbol{Y})] = \int g\big((\nabla_\theta \log f_1(\theta_1,x_1)^T,\ldots,\nabla_\theta \log f_N(\theta_N,x_N)^T)^T\big)\;\; f_*(\boldsymbol{x})\,d\boldsymbol{x}$$

for every measurable function $g:\mathbb{R}^{Nd}\to\mathbb{R}_+$. The associated mean field given by Equation (8) 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 Theorem 1 or Theorem 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 [23], [24], [25]). 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 $\xrightarrow{\mathcal{D}}$ 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 $f(\theta_\star,\cdot)$ [22, Chap. 6]. We now turn to the distributed algorithm and to that end, we apply Theorems 4 and 5. 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 5 shows that on the

set $\{\lim_n \boldsymbol{\theta}_n = \mathbf{1} \otimes \theta_\star\}$, the averaged sequence $\bar{\boldsymbol{\theta}}_n$ satisfies $\sqrt{n}(\bar{\boldsymbol{\theta}}_n - \mathbf{1} \otimes \theta_\star) \xrightarrow{\mathcal{D}} \mathbf{1} \otimes Z$ 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 non averaged algorithm. In order to make a fair comparison with the centralized ML algorithm, we restrict the use of Theorem 4 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 4 shows that on the set $\{\lim_n \boldsymbol{\theta}_n = \mathbf{1} \otimes \theta_\star\}$, the sequence of estimates $\boldsymbol{\theta}_n$ satisfies $\sqrt{n}(\boldsymbol{\theta}_n - \mathbf{1} \otimes \theta_\star) \xrightarrow{\mathcal{D}} \mathbf{1} \otimes Z$ 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 non averaged algorithm.

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 Figure 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 [26] 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 $c_1/n^{0.6}$ for $n \leq 10000$ iterations, $c_2(\log n/n)^{0.6}$ for $10000 < n \leq 20000$ and $c_3(\log n/n)^{0.6}$ for $n > 20000$ with $c_1 < c_2 < c_3$. 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 Figure 2. Four paths (starting from the same value $\boldsymbol{\theta}_0$) are run and we display $n \mapsto (1/N)|\boldsymbol{\theta}_n - \mathbf{1} \otimes \theta_\star|$ for $n \leq 50000$. Note the role of the step size sequence in the rate of convergence (compare the definition of γ_n above and the changes in the slopes at time $n = 10000$ and $n = 20000$).

VI. PROOFS

A. Notations

For a positive deterministic sequence $(a_n)_{n\geq 1}$, $o(a_n)$ stands for 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 $\|X\|_p := \mathbb{E}(|X|^p)^{1/p}$. $o_{L^p}(a_n)$ stands for any $\mathbb{R}^\ell$-valued r.v. $(X_n)_{n\geq 1}$ such that $\lim_{n\to\infty} a_n^{-1}\|X_n\|_p = 0$; $\mathcal{O}_{L^p}(a_n)$ stands for any $\mathbb{R}^\ell$-valued r.v. $(X_n)_{n\geq 1}$ such that $\limsup_n a_n^{-1}\|X_n\|_p < \infty$; and $\mathcal{O}_{w.p.1.}(a_n)$ 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.

We start with a preliminary lemma which will be crucial for most of the proofs.

B. Preliminary result

Lemma 1 (Agreement): Under Assumptions **1a-b)**, **2**, **3a-c**), **4a**) and **5**,

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) $\sup_{n\geq 1} \mathbb{E}V(\langle\boldsymbol{\theta}_n\rangle) < \infty$,

where $\langle\boldsymbol{x}\rangle$ and $\boldsymbol{x}_\perp$ are given by (6) and (7).

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 exist a constant $M > 0$ and an integer n_0 such that for any $n \geq n_0$:

$$u_n \leq \rho_n u_{n-1} + \gamma_n M\sqrt{u_{n-1}}(1 + u_{n-1} + v_{n-1})^{1/2} + \gamma_n^2 M\left(1 + u_{n-1} + v_{n-1}\right) , \tag{16}$$

$$v_n \leq v_{n-1} + M u_{n-1} + \gamma_n M\sqrt{u_{n-1}}\left(1 + u_{n-1} + v_{n-1}\right)^{1/2} + \gamma_n^2 M(1 + u_{n-1} + v_{n-1}) . \tag{17}$$

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 (29) and (30).

Proof of (16). 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:

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

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 **4a)**, we have the following estimate $\mathbb{E}[|\boldsymbol{Y}_n|^2] \leq C_2\,(1 + v_{n-1} + u_{n-1})$. This completes the proof of (16), for any constant M larger than $1 + C_2$.

Proof of (17). We use the following Taylor-Lagrange expansion of the Lyapunov function V at point $\langle \boldsymbol{\theta}_n \rangle$. There exists $\hat{\theta}_n \in \mathbb{R}^d$ such that $|\hat{\theta}_n - \langle \boldsymbol{\theta}_{n-1} \rangle| \leq |\langle \boldsymbol{\theta}_n \rangle - \langle \boldsymbol{\theta}_{n-1} \rangle|$ and

$$V(\langle \boldsymbol{\theta}_n \rangle) = V(\langle \boldsymbol{\theta}_{n-1} \rangle) + \nabla V(\hat{\theta}_n)^T (\langle \boldsymbol{\theta}_n \rangle - \langle \boldsymbol{\theta}_{n-1} \rangle) \ .$$

Under Assumption **3a)**, ∇V is a Lipschitz function. Thus, $|\nabla V(\hat{\theta}_n) - \nabla V(\langle \boldsymbol{\theta}_{n-1} \rangle)| \leq K_{Lip} |\langle \boldsymbol{\theta}_n \rangle - \langle \boldsymbol{\theta}_{n-1} \rangle|$, where K_{Lip} denotes the Lipschitz constant. Therefore,

$$V(\langle \boldsymbol{\theta}_n \rangle) \leq V(\langle \boldsymbol{\theta}_{n-1} \rangle) + \nabla V(\langle \boldsymbol{\theta}_{n-1} \rangle)^T (\langle \boldsymbol{\theta}_n \rangle - \langle \boldsymbol{\theta}_{n-1} \rangle) + K_{Lip} |\langle \boldsymbol{\theta}_n \rangle - \langle \boldsymbol{\theta}_{n-1} \rangle|^2 \ . \tag{18}$$

We need to evaluate the difference $\langle \boldsymbol{\theta}_n \rangle - \langle \boldsymbol{\theta}_{n-1} \rangle$. By (2),

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

Therefore,

$$\begin{aligned} \langle \boldsymbol{\theta}_n \rangle - \langle \boldsymbol{\theta}_{n-1} \rangle &= \left(\frac{\mathbf{1}^T W_n - \mathbf{1}^T}{N} \otimes I_d\right) \boldsymbol{\theta}_{n-1} + \left(\frac{\mathbf{1}^T W_n}{N} \otimes I_d\right) \gamma_n \boldsymbol{Y}_n \\ &= \left(\frac{\mathbf{1}^T W_n - \mathbf{1}^T}{N} \otimes I_d\right) \boldsymbol{\theta}_{\perp,n-1} + \left(\frac{\mathbf{1}^T W_n}{N} \otimes I_d\right) \gamma_n \boldsymbol{Y}_n \ , \end{aligned} \tag{19}$$

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)**:

$$\mathbb{E}[\langle \boldsymbol{\theta}_n \rangle - \langle \boldsymbol{\theta}_{n-1} \rangle | \mathcal{F}_{n-1}] = \gamma_n \mathbb{E}_{\boldsymbol{\theta}_{n-1}} \langle \boldsymbol{Y}_n \rangle \ . \tag{20}$$

Plugging (20) into (18),

$$\mathbb{E}[V(\langle \boldsymbol{\theta}_n \rangle) | \mathcal{F}_{n-1}] \leq V(\langle \boldsymbol{\theta}_{n-1} \rangle) + \gamma_n \nabla V(\langle \boldsymbol{\theta}_{n-1} \rangle)^T \mathbb{E}_{\boldsymbol{\theta}_{n-1}} \langle \boldsymbol{Y}_n \rangle + K_{Lip} \mathbb{E}[|\langle \boldsymbol{\theta}_n \rangle - \langle \boldsymbol{\theta}_{n-1} \rangle|^2 | \mathcal{F}_{n-1}] \ .$$

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,

$$\mathbb{E}[V(\langle \boldsymbol{\theta}_n\rangle)|\mathcal{F}_{n-1}] \leq V(\langle \boldsymbol{\theta}_{n-1}\rangle) + \gamma_n \nabla V(\langle \boldsymbol{\theta}_{n-1}\rangle)^T (\mathbb{E}_{\boldsymbol{\theta}_{n-1}}\langle \boldsymbol{Y}_n\rangle - h(\langle \boldsymbol{\theta}_{n-1}\rangle)) + K_{Lip}\mathbb{E}[|\langle \boldsymbol{\theta}_n\rangle - \langle \boldsymbol{\theta}_{n-1}\rangle|^2|\mathcal{F}_{n-1}] \ .$$

Using successively the conditions **4a)** and **3c)**, we have the estimate

$$\begin{aligned} \nabla V(\langle \boldsymbol{\theta}_{n-1}\rangle)^T (\mathbb{E}_{\boldsymbol{\theta}_{n-1}}\langle \boldsymbol{Y}_n\rangle - h(\langle \boldsymbol{\theta}_{n-1}\rangle)) &\leq |\nabla V(\langle \boldsymbol{\theta}_{n-1}\rangle)|\, C_2 |\boldsymbol{\theta}_{\perp,n-1}| \\ &\leq \sqrt{C_1} C_2 \sqrt{1 + V(\langle \boldsymbol{\theta}_{n-1}\rangle)}\, |\boldsymbol{\theta}_{\perp,n-1}| \ . \end{aligned}$$

Using Cauchy-Schwartz inequality, the expectation of the above quantity is no larger than $\sqrt{C_1}C_2\sqrt{u_{n-1}(1+v_{n-1})}$. We obtain:

$$v_n \leq v_{n-1} + \gamma_n \sqrt{C_1} C_2 \sqrt{u_{n-1}(1+u_{n-1}+v_{n-1})} + K_{Lip}\mathbb{E}[|\langle \boldsymbol{\theta}_n\rangle - \langle \boldsymbol{\theta}_{n-1}\rangle|^2] \ , \tag{21}$$

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 (19),

$$\mathbb{E}[|\langle \boldsymbol{\theta}_n\rangle - \langle \boldsymbol{\theta}_{n-1}\rangle|^2]^{1/2} \leq \mathbb{E}\left[\left|\left(\frac{\mathbf{1}^T W_n - \mathbf{1}^T}{N} \otimes I_d\right)\boldsymbol{\theta}_{\perp,n-1}\right|^2\right]^{1/2} + \mathbb{E}\left[\left|\left(\frac{\mathbf{1}^T W_n}{N} \otimes I_d\right)\gamma_n \boldsymbol{Y}_n\right|^2\right]^{1/2} \tag{22}$$

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:

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

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

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

where we used assumption **4a)**. Putting this together with (22),

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

where $C > 0$ is some constant chosen large enough. Plugging the above inequality into (21),

$$\begin{aligned}v_n \leq v_{n-1} + (K_{Lip}C)u_{n-1} + (\sqrt{C_1}C_2 + K_{Lip}C)\gamma_n\sqrt{u_{n-1}(1+u_{n-1}+v_{n-1})} \\ + K_{Lip}C\gamma_n^2\,(1+u_{n-1}+v_{n-1})\ .\end{aligned}$$

This proves that (17) holds for any M chosen large enough. ■

Corollary 1 (of Lemma 1): Under the assumptions of Lemma 1, $\sup_n \mathbb{E}\,[|\boldsymbol{Y}_n|^2] < \infty$.

Proof: By Assumptions **1b)** and **4a)**:

$$\mathbb{E}\left[|\boldsymbol{Y}_n|^2\right] = \mathbb{E}\left[\mathbb{E}_{\boldsymbol{\theta}_{n-1}}\left[|\boldsymbol{Y}|^2\right]\right] \leq C_2\left(1 + \mathbb{E}\,[V(\langle\boldsymbol{\theta}_{n-1}\rangle)] + \mathbb{E}\left[|\boldsymbol{\theta}_{\perp,n-1}|^2\right]\right)\ . \tag{23}$$

The proof is concluded by Lemma 1. ■

C. 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. By Lemma 1, $(\boldsymbol{\theta}_{\perp,n})_{n\geq 1}$ converges to zero w.p.1 and in L^2. Therefore, 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 [27, Theorems 2.2. and 2.3.].

Lemma 2: Under Assumptions **1**a-b), **2**, **3a-c**), **4a**) and **5**, 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 $\sup_n|\sum_{k=1}^n \gamma_k\zeta_k| < \infty$ almost-surely.

Proof: Eqs. (2) and (6) along with assumption **1a)** yield:

$$\langle\boldsymbol{\theta}_n\rangle = \langle\boldsymbol{\theta}_{n-1}\rangle + \gamma_n\langle\boldsymbol{Z}_n\rangle\ , \qquad \text{where} \qquad \boldsymbol{Z}_n := (W_n\otimes I_d)(\boldsymbol{Y}_n + \gamma_n^{-1}\boldsymbol{\theta}_{\perp,n-1}) \tag{24}$$

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

$$\begin{aligned}e_n &:= \langle(W_n\otimes I_d)(\boldsymbol{Y}_n + \gamma_n^{-1}\boldsymbol{\theta}_{\perp,n-1})\rangle - \mathbb{E}_{\boldsymbol{\theta}_{n-1}}[\langle\boldsymbol{Y}\rangle] \\ \xi_n &:= \mathbb{E}_{\boldsymbol{\theta}_{n-1}}[\langle\boldsymbol{Y}\rangle] - \mathbb{E}_{\mathbf{1}\otimes\langle\boldsymbol{\theta}_{n-1}\rangle}[\langle\boldsymbol{Y}\rangle]\ .\end{aligned}$$

By Assumption **4a)** and the inequality $2ab \leq a^2 + b^2$, there exists a constant C such that

$$\mathbb{E}\left|\sum_{n\geq 1}\gamma_n\xi_n\right| \leq C\ \left(\sum_{n\geq 1}\gamma_n^2 + \sum_{n\geq 1}\mathbb{E}\left|\boldsymbol{\theta}_{\perp,n-1}\right|^2\right)\ . \tag{25}$$

Therefore, the RHS in (25) is finite under the condition **2** and Lemma 1, thus implying that $\sum_{n\geq 1}\gamma_n\xi_n$ converges w.p.1.

Since $\mathbb{E}\left[e_n\left|\mathcal{F}_{n-1}\right.\right] = 0$, the sequence $\left(S_n := \sum_{k=1}^n \gamma_k e_k\right)_{n\geq 1}$ is a martingale. We prove that it converges almost surely by estimating its second order moment. For any $k \geq 1$,

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

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:

$$\mathbb{E}\left[|S_n|^2\right] \leq C\sum_{n\geq 1}\gamma_n^2\,\mathbb{E}\left[\left|\boldsymbol{Y}_n + \gamma_n^{-1}\boldsymbol{\theta}_{\perp,n-1}\right|^2\right] \leq 2C\sum_{n\geq 1}\gamma_n^2\,\mathbb{E}\left[|\boldsymbol{Y}_n|^2\right] + 2C\sum_{n\geq 1}\mathbb{E}\left[|\boldsymbol{\theta}_{\perp,n-1}|^2\right]\ .$$

By Lemma 1, Corollary 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. [28, Corollary 2.2.]). This concludes the proof.

■

D. Proof of Theorem 3

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

$$\Phi_{n,k} := (V_n \otimes I_d)(V_{n-1} \otimes I_d)\cdots(V_k \otimes I_d)\ . \tag{26}$$

Note that by Assumptions **1b-c)**,

$$\begin{aligned}\|\Phi_{n,k}X\|_2^2 &= \mathbb{E}[X^T\Phi_{n-1,k}^T(V_n^T V_n \otimes I_d)\Phi_{n-1,k}X] = \mathbb{E}[X^T\Phi_{n-1,k}^T\mathbb{E}(V_n^T V_n \otimes I_d)\Phi_{n-1,k}X] \\ &\leq \rho\ \mathbb{E}[X^T\Phi_{n-1,k}^T\Phi_{n-1,k}X]\ = \rho\|\Phi_{n-1,k}X\|_2^2\ .\end{aligned} \tag{27}$$

From (2) 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 $n \geq 1$, $\boldsymbol{\theta}_{\perp,n} = (V_n \otimes I_d)(\boldsymbol{\theta}_{\perp,n-1} + \gamma_n\boldsymbol{Y}_{\perp,n})$. By induction,

$$\boldsymbol{\theta}_{\perp,n} = \sum_{k=1}^n \gamma_k\Phi_{n,k}\boldsymbol{Y}_{\perp,k} + \Phi_{n,1}\boldsymbol{\theta}_{\perp,0} \tag{28}$$

where $\Phi_{n,k}$ is defined by (26). By (27) and Assumption **1c)**, the second term in the RHS of (28) is a $\mathcal{O}_{L^2}(\rho^{n/2})$. We now consider the first term in the RHS of (28). Using Minkowski's inequality and Equation (27)

$$\|\sum_{k=1}^{n}\gamma_k\Phi_{n,k}\boldsymbol{Y}_{\perp,k}\|_2 \leq \sum_{k=1}^{n}\gamma_k\|\Phi_{n,k}\boldsymbol{Y}_{\perp,k}\|_2 \leq \sum_{k=1}^{n}\gamma_k\sqrt{\rho}^{n-k+1}\|\boldsymbol{Y}_{\perp,k}\|_2 \ .$$

By [29, Result 178,pp.38], the RHS is upper bounded by $C\ \rho(1-\sqrt{\rho})^{-1}$ with $C := \limsup_{n\to\infty}\|\boldsymbol{Y}_{\perp,n}\|_2$, which is finite by Corollary 1. This concludes the proof.

E. Proof of Theorem 4

Assumption 2 implies that $\lim_n \rho^{n/2}\gamma_n^{-2} = 0$. Therefore, by Theorem 3, the sequence of r.v. $(\gamma_n^{-1/2}\boldsymbol{\theta}_{\perp,n})_n$ converges in probability to zero. 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. $(\gamma_n^{-1/2}(\langle\boldsymbol{\theta}_n\rangle - \theta_\star))_{n\geq 0}$ converges in distribution to Z (under the conditional distribution given the event $\{\lim_q \theta_q = \mathbf{1}\otimes\theta_\star\}$ which, under Lemma 1 is the same as the conditional distribution given the event $\{\lim_q\langle\theta_q\rangle = \theta_\star\}$). To that goal, we write

$$\langle\boldsymbol{\theta}_n\rangle - \theta_\star = \langle\boldsymbol{\theta}_{n-1}\rangle - \theta_\star + \gamma_n h\left(\langle\boldsymbol{\theta}_{n-1}\rangle\right) + \gamma_n e_n\mathbf{1}_{|\boldsymbol{\theta}_{n-1}-\theta_\star|\leq\delta} + \gamma_n\xi_n + \gamma_n e_n\mathbf{1}_{|\boldsymbol{\theta}_{n-1}-\theta_\star|>\delta}$$

where $\xi_n := \mathbb{E}_{\boldsymbol{\theta}_{n-1}}[\langle\boldsymbol{Y}\rangle] - \mathbb{E}_{\mathbf{1}\otimes\langle\boldsymbol{\theta}_{n-1}\rangle}[\langle\boldsymbol{Y}\rangle]$ and

$$e_n := \langle(W_n\otimes I_d)(\boldsymbol{Y}_n + \gamma_n^{-1}\boldsymbol{\theta}_{\perp,n-1})\rangle - \mathbb{E}_{\boldsymbol{\theta}_{n-1}}[\langle\boldsymbol{Y}\rangle] = \langle\boldsymbol{Y}_n\rangle - \mathbb{E}_{\boldsymbol{\theta}_{n-1}}[\langle\boldsymbol{Y}\rangle] \ ,$$

since $\mathbf{1}^T W_n = \mathbf{1}^T$. We then check the conditions C1 to C4 of [20, Theorem 1] (see also [30, Theorem 1]). Under the assumptions **6** and **8a)**, the conditions C1 and C4 of [20, Theorem 1] are satisfied. We now prove C2b: there exists a constant C such that

$$\begin{aligned}\mathbb{E}\left[|e_{n+1}|^{2+\tau}\mathbf{1}_{|\boldsymbol{\theta}_n-\mathbf{1}\otimes\theta_\star|\leq\delta}\right] &\leq C\ \mathbb{E}\left[|\mathbb{E}_{\boldsymbol{\theta}_n}\left[\langle\boldsymbol{Y}\rangle\right]|^{2+\tau}\mathbf{1}_{|\boldsymbol{\theta}_n-\mathbf{1}\otimes\theta_\star|\leq\delta}\right] + C\ \mathbb{E}\left[|\langle\boldsymbol{Y}_{n+1}\rangle|^{2+\tau}\mathbf{1}_{|\boldsymbol{\theta}_n-\mathbf{1}\otimes\theta_\star|\leq\delta}\right] \\ &\leq 2C\sup_{|\boldsymbol{\theta}-\mathbf{1}\otimes\theta_\star|\leq\delta}\mathbb{E}_{\boldsymbol{\theta}}\left[|\langle\boldsymbol{Y}\rangle|^{2+\tau}\right]\end{aligned}$$

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

$$\mathbb{E}\left[e_{n+1}e_{n+1}^T|\mathcal{F}_n\right]\mathbf{1}_{|\boldsymbol{\theta}_n-\mathbf{1}\otimes\theta_\star|\leq\delta} = \left\{\mathbb{E}_{\boldsymbol{\theta}_n}\left[\langle\boldsymbol{Y}\rangle\langle\boldsymbol{Y}\rangle^T\right] - \mathbb{E}_{\boldsymbol{\theta}_n}\left[\langle\boldsymbol{Y}\rangle\right]\left(\mathbb{E}_{\boldsymbol{\theta}_n}\left[\langle\boldsymbol{Y}\rangle\right]\right)^T\right\}\mathbf{1}_{|\boldsymbol{\theta}_n-\mathbf{1}\otimes\theta_\star|\leq\delta}.$$

By Assumptions **4** and **7**, this term converges w.p.1 to Υ on the set $\{\lim_k\boldsymbol{\theta}_k = \mathbf{1}\otimes\theta_\star\}$ and since $\{\lim_k\boldsymbol{\theta}_k = \mathbf{1}\otimes\theta_\star\}$ = $\{\lim_k\langle\boldsymbol{\theta}_k\rangle = \theta_\star\}$ w.p.1 (as a consequence of Lemma 1), it also converges w.p.1 to Υ on the set $\{\lim_k\langle\boldsymbol{\theta}_k\rangle = \theta_\star\}$. This concludes the proof of C2.

We now consider the condition C3 of [20] with $r_n = \xi_n + e_n \mathbf{1}_{|\boldsymbol{\theta}_{n-1} - \mathbf{1}\otimes\theta_\star|>\delta}$. By Assumption **4a)**, Theorem 3 and Lemma 1, there exists a constant C such that

$$\gamma_n^{-1/2}\mathbb{E}\left[|\xi_n|\mathbf{1}_{\lim_k\langle\boldsymbol{\theta}_k\rangle=\theta_\star}\right] = \gamma_n^{-1/2}\mathbb{E}\left[|\xi_n|\mathbf{1}_{\lim_k \boldsymbol{\theta}_k=\mathbf{1}\otimes\theta_\star}\right] \leq C\ \left(\gamma_n^{-1}\mathbb{E}\left[|\boldsymbol{\theta}_{\perp,n}|^2\right]\right)^{1/2}$$

and the RHS tends to zero as $n \to \infty$. On the set $\{\lim_n\langle\boldsymbol{\theta}_n\rangle = \theta_\star\}$ (which, as discussed above, is equal w.p.1 to 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 so that $\gamma_n\sum_{k=1}^n e_k\mathbf{1}_{|\boldsymbol{\theta}_{k-1}-\mathbf{1}\otimes\theta_\star|>\delta}$ is $\mathcal{O}_{w.p.1}o_{L^1}(1)$. This concludes the proof of the condition C3 of [20], and the proof of Theorem 4.

F. Proof of Theorem 5

We preface the proof by a preliminary result, established by [20, Theorem 2] (see also [19] for a similar result obtained under stronger assumptions).

Theorem 6: Let $(\gamma_n)_n$ be a deterministic positive sequence such that $\log(\gamma_k/\gamma_{k+1}) = o(\gamma_k)$ and satisfying Assumption **8b-c)**. Consider the random sequence $(u_n)_n$ given by

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

where

AVER 1:

(a) $u_\star$ is a zero of the mean field: $h(u_\star) = 0$.

(b) 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.

AVER 2:

(a) $(e_n)_{n\geq 1}$ is a $\mathcal{F}_n$-adapted martingale-increment sequence.

(b) There exist $\tau > 0$ and $\delta \in (0, +\infty]$ s.t. $\sup_k \mathbb{E}\left[|e_k|^{2+\tau}\mathbf{1}_{|u_{k-1}-u_\star|\leq\delta}\right] < \infty$.

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

AVER 3: $(\xi_n)_{n\geq 1}$ is a $\mathcal{F}_n$-adapted sequence s.t.

(a) $\gamma_n^{-1/2}\ |\xi_n|\mathbf{1}_{\lim_q u_q=u_\star} = \ \mathcal{O}_{w.p.1}(1)\mathcal{O}_{L^2}(1)$

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

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

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

Proof of Theorem 5. By Theorem 3 and Assumption **8c)**, $\sqrt{N}^{-1}\sum_{n=1}^N \boldsymbol{\theta}_{\perp,n}$ converges in L^2 to zero. 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 AVER1 to AVER3 of Theorem 6 with $u_n = \langle\boldsymbol{\theta}_n\rangle$ and e_n, ξ_n defined as in the proof of Theorem 4. Under Assumption 6, AVER1 holds. AVER2 is proved along the same lines as in the proof of Theorem 4. Finally, by Assumption **4a)** and Theorem 3, $\gamma_n^{-1}\mathbb{E}\left[|\xi_n|^2\mathbf{1}_{\lim_k \boldsymbol{\theta}_k=\mathbf{1}\otimes\theta_\star}\right] = \mathcal{O}(\gamma_n)$; and

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

The RHS tends to zero under Assumption **8c)** thus showing AVER3.

APPENDIX

Lemma 3: Let $(\gamma_n)_{n\ge0}$, $(\rho_n)_{n\ge0}$ 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 (16) and (17) hold true for $n \ge n_0$, and $u_{n_0} + v_{n_0} < \infty$. Then: *i)* $\sup_n v_n < \infty$, *ii)* $\limsup_n \phi_n u_n < \infty$ for any positive sequence $(\phi_n)_{n\ge0}$ such that

$$\limsup_n\left(\gamma_n\sqrt{\phi_n} + \frac{\phi_{n-1}}{\phi_n}\right) < \infty\ , \quad \liminf_n(\gamma_n\sqrt{\phi_n})^{-1}\left(\frac{\phi_{n-1}}{\phi_n} - \rho_n\right) > 0\ , \tag{29}$$

$$\sum_n \phi_n^{-1} < \infty\ . \tag{30}$$

Remark 2: If the sequences $(\gamma_n, \rho_n)_{n\ge0}$ are such that

$$\limsup_n\left(\frac{\gamma_n}{\gamma_{n-1}} + \frac{1-\rho_{n-1}}{1-\rho_n}\right) < \infty\ , \quad \liminf_n \frac{1}{1-\rho_n}\left(\frac{(1-\rho_{n-1})^2}{(1-\rho_n)^2}\frac{\gamma_n^2}{\gamma_{n-1}^2} - \rho_n\right) > 0\ , \tag{31}$$

$$\sum_n \gamma_n^2(1-\rho_n)^{-2} < \infty\ , \tag{32}$$

then the conditions (29) and (30) 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 $0 \le \eta < 1 \wedge (\xi - 1/2)$.

Proof: • 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 $n \geq n_0 + 1$:

$$a_n = \rho_n a_{n-1} + \tilde{\gamma}_n \sqrt{a_{n-1}}\,(1 + a_{n-1} + b_{n-1})^{1/2} + \tilde{\gamma}_n^2(1 + a_{n-1} + b_{n-1}) \tag{33}$$

$$b_n = b_{n-1} + M a_{n-1} + \tilde{\gamma}_n \sqrt{a_{n-1}}(1 + a_{n-1} + b_{n-1})^{1/2} + \tilde{\gamma}_n^2(1 + a_{n-1} + b_{n-1})\ . \tag{34}$$

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, Eq. (33) implies:

$$a_n \leq \rho_n a_{n-1} + \tilde{\gamma}_n \sqrt{a_{n-1}}\,(1 + 2\,A_{n-1})^{1/2} + \tilde{\gamma}_n^2(1 + 2A_{n-1})\ . \tag{35}$$

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

$$\frac{a_n}{A_n} \leq \rho_n \frac{a_{n-1}}{A_{n-1}} + \tilde{\gamma}_n \sqrt{\frac{a_{n-1}}{A_{n-1}}} \left(\frac{1}{A_{n_0}} + 2\right)^{1/2} + \tilde{\gamma}_n^2 \left(\frac{1}{A_{n_0}} + 2\right)\ . \tag{36}$$

• Define $L^2 := 1/A_{n_0} + 2$, and $c_n := \phi_n a_n / A_n$. By (36), for any $n \geq n_0 + 1$,

$$c_n \leq \rho_n \frac{\phi_n}{\phi_{n-1}} c_{n-1} + L\tilde{\gamma}_n \sqrt{c_{n-1}\phi_n} \sqrt{\frac{\phi_n}{\phi_{n-1}}} + L^2\ \tilde{\gamma}_n^2 \phi_n, \tag{37}$$

and under the assumption (29), there exist $n_1 \geq n_0$ and a constant $\xi > 0$ such that for any $n \geq n_1$,

$$\sqrt{\frac{\phi_{n-1}}{\phi_n}} L\xi \left\{1 + \xi L \tilde{\gamma}_n \sqrt{\phi_{n-1}}\right\} \leq \left(\frac{\phi_{n-1}}{\phi_n} - \rho_n\right) \left(\tilde{\gamma}_n \sqrt{\phi_n}\right)^{-1}\ . \tag{38}$$

Define

$$A := \max\left(\frac{1}{\xi}, \frac{1}{\xi^2}, c_{n_1}\right)\ . \tag{39}$$

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 (37) and (39), 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 (38), 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.

• 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 Eq. (35), 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 (29), $\limsup_n\{\tilde{\gamma}_n/\sqrt{\phi_n}\}\phi_n < \infty$). By assumptions, $\sum_n\{\phi_{n-1}^{-1} + \tilde{\gamma}_n^2\} < \infty$, $(A_n)_{n\geq n_0}$ is therefore bounded.

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

■

REFERENCES

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

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

[3] V.D. Blondel, J.M. Hendrickx, A. Olshevsky, and J.N. Tsitsiklis, "Convergence in multiagent coordination, consensus, and flocking," in *Decision and Control, 2005 and 2005 European Control Conference. CDC-ECC '05. 44th IEEE Conference on*, dec. 2005, pp. 2996 – 3000.

[4] C. Lopes and A.H. Sayed, "Distributed processing over adaptive networks," in *Adaptive Sensor Array Processing Workshop*, June 2006, pp. 1–5.

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

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

[7] A. Nedic, A. Ozdaglar, and P.A. Parrilo, "Constrained Consensus and Optimization in Multi-Agent Networks," *IEEE Transactions on Automatic Control*, vol. 55, no. 4, pp. 922–938, April 2010.

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

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

[10] J. Tsitsiklis, D. Bertsekas, and M. Athans, "Distributed asynchronous deterministic and stochastic gradient optimization algorithms," *Automatic Control, IEEE Transactions on*, vol. 31, no. 9, pp. 803 – 812, sep 1986.

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

[12] S. Sundhar Ram, A. Nedic, and V. Veeravalli, "Distributed stochastic subgradient projection algorithms for convex optimization," *Journal of Optimization Theory and Applications*, vol. 147, pp. 516–545, 2010, 10.1007/s10957-010-9737-7.

[13] A. Nedic and A. Ozdaglar, "Distributed Subgradient Methods for Multi-Agent Optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.

[14] P. Bianchi, G. Fort, W. Hachem, and J. Jakubowicz, "Performance of a Distributed Robbins-Monro Algorithm for Sensor Networks," in *EUSIPCO*, Barcelona, Spain, 2011.

[15] T.C. Aysal, M.E. Yildiz, A.D. Sarwate, and A. Scaglione, "Broadcast Gossip Algorithms for Consensus," *IEEE Transactions on Signal Processing*, vol. 57, no. 7, pp. 2748–2761, 2009.

[16] B. Polyak, "New stochastic approximation type procedures," *Automation and remote control*, vol. 51, pp. 98–107, 1990.

[17] F. Bénézit, *Distributed Average Consensus for Wireless Sensor Networks*, Ph.D. thesis, École Polytechnique Fédérale de Lausanne, 2009.

[18] P. Bianchi and J. Jakubowicz, "On the convergence of a multi-agent projected stochastic gradient algorithm," *Submitted*, 2011, [online] arXiv:1107.2526v1.

[19] B. Delyon, "Stochastic Approximation with Decreasing Gain: Convergence and Asymptotic Theory," *Unpublished Lecture Notes, http://perso.univ-rennes1.fr/bernard.delyon/as_cours.ps*, 2000.

[20] G. Fort, "A Central Limit Theorem for a stochastic approximation algorithm and its Polyak-averaged version," Tech. Rep., Preprint, 2012, [online] http://perso.telecom-paristech.fr/∼gfort/Preprints/CLTforSA.pdf.

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

[22] E. L. Lehmann and George Casella, *Theory of point estimation*, Springer Texts in Statistics. Springer-Verlag, New York, second edition, 1998.

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

[24] Michel Benaim, "Dynamics of stochastic approximation algorithms," in *Séminaire de Probabilités, XXXIII*, vol. 1709 of *Lecture Notes in Math.*, pp. 1–68. Springer, Berlin, 1999.

[25] Hai-Tao Fang and Han-Fu Chen, "Stability and instability of limit points for stochastic approximation algorithms," *Automatic Control, IEEE Transactions on*, vol. 45, no. 3, pp. 413 –420, mar 2000.

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

[27] 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.

[28] P. Hall and C. C. Heyde, *Martingale Limit Theory and its Application*, Academic Press, New York, London, 1980.

[29] G. Pólya and G. Szegő, *Problems and theorems in analysis: Series. Integral calculus. Theory of functions*, Springer (Classics in mathematics), 1998.

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

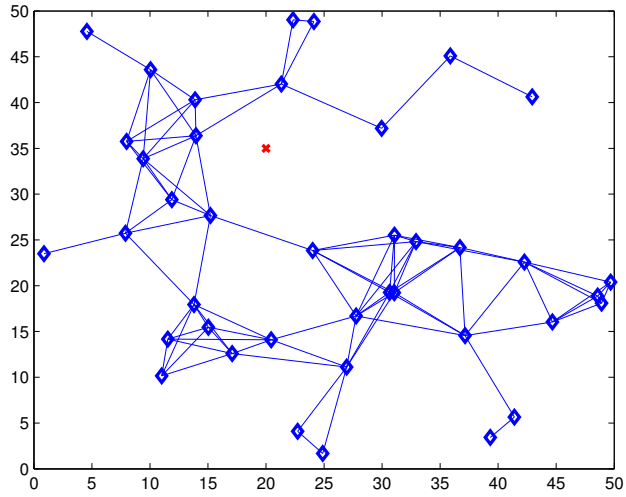

Fig. 1. $N = 40$ sensors (diamonds) with the graph (line segments) and the source (star)

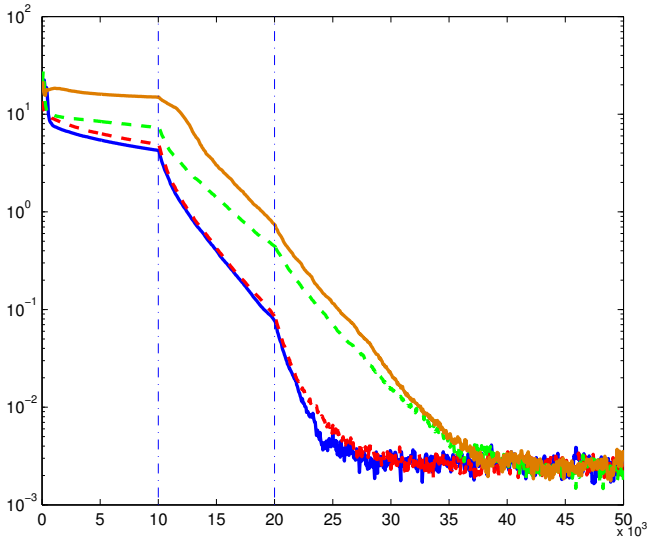

Fig. 2. Cumulative relative error (over the N sensors) when estimating θ_*, as a function of the number of iterations.