Success and Failure of Adaptation-Diffusion Algorithms for Consensus in Multi-Agent Networks

Gemma Morral*, Pascal Bianchi and Gersende Fort

***Abstract*—This paper investigates the problem of distributed stochastic approximation in multi-agent systems. The algorithm under study consists of two steps: a local stochastic approximation step and a gossip step which drives the network to a consensus. The gossip step uses row-stochastic matrices to weight network exchanges.**

We first prove the convergence of a distributed optimization algorithm, when the function to optimize may not be convex and the communication protocol in independent of the observations. In that case, we prove that the average estimate converges to a consensus; we also show that the set of limit points is not necessarily the set of the critical points of the function to optimize and is affected by the Perron eigenvector of a mean-matrix describing the communication protocol. Discussion about the success or failure of convergence to the minimizers of the function to optimize is also addressed. In a second part of the paper, we extend the convergence results to the more general context of distributed stochastic approximation.

I. INTRODUCTION

Distributed stochastic approximation has been recently proposed using different cooperative approaches. In the so-called *incremental* approach (see for instance [1], [2]) a message containing an estimate of the quantity of interest iteratively travels all over the network. This paper focuses on another cooperative approach based on *average consensus* techniques where the estimates computed locally by each agent are combined through the network. This idea traces back to [3] where a network of processors seeks to optimize some objective function *known* by all agents (possibly up to some additive noise).

In this paper, we consider the following cooperative approach. Let a network be composed by N agents, or nodes. Agents seek to find a consensus on some global parameter by means of local observations and peer-to-peer communications. Node i ($i = 1, \ldots, N$) generates a $\mathbb{R}^d$-valued stochastic process $(\theta_{n,i})_{n\geq 0}$, initialized at some arbitrary $\theta_{0,i} \in \mathbb{R}^d$. Let $(\gamma_n)_{n\geq 1}$ be a deterministic positive step size sequence. At time n, for all $i = 1, \ldots, N$:

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

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

where the $\mathbb{R}^d$-valued random process $(Y_{n,i})_{n\geq 0}$ represents the observations made by agent i.

`[Gossip step]` Node i is able to observe the values $\tilde{\theta}_{n,j}$ of some other nodes 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 non-negative random coefficients such that $\sum_{j=1}^{N} w_n(i,j) = 1$ for any i. The random matrix $W_n := [w_n(i,j)]_{i,j=1}^{N}$ represents the network connections between the nodes at time n. One simply set $w_n(i,j) = 0$ whenever nodes i and j are unable to communicate at time n.

Application to distributed optimization. In many applications related to machine learning and sensor networks (we refer to [8] and [10]–[12] for more details) or smart grids [13], one seeks to minimize a sum of local private cost functions f_i of the agents:

$$\min_{\theta \in \mathbb{R}^d} \ \sum_{i=1}^{N} f_i(\theta) \ . \tag{3}$$

In this context, the distributed Algorithm (1)-(2) reduces to a distributed stochastic gradient algorithm by letting

$$Y_{n,i} = -\nabla f_i(\theta_{n-1,i}) + \xi_{n,i} \tag{4}$$

where ∇ is the gradient operator and $\xi_{n,i}$ represents some random perturbation which possibly occurs when observing the gradient. The function f_i is supposed to be unknown from the other agents j, $j \neq i$. In this paper, we handle the case where functions f_i are not necessarily convex. Of course, in that case, there is generally no hope to ensure the convergence to a minimizer to (3). Instead, a more realistic objective is to achieve *critical points* of the objective function *i.e.*, points θ such that $\sum_i \nabla f_i(\theta) = 0$.

Doubly and non-doubly stochastic matrices. In most works (see for instance [14], [15]), the matrices $(W_n)_{n\geq 1}$ are assumed *doubly stochastic*, meaning that $W_n^T \mathbf{1} = W_n \mathbf{1} = \mathbf{1}$ where $\mathbf{1}$ is the $N \times 1$ vector whose components are all equal to one and where ^{T} denotes transposition. Although row-stochasticity ($W_n \mathbf{1} = \mathbf{1}$) is rather easy to ensure in practice, column-stochasticity ($W_n^T \mathbf{1} = \mathbf{1}$) implies more stringent restrictions on the communication protocol. For instance, in [16], each one-way transmission from an agent i to another agent j requires at the same time a feedback link from j to i. As a matter of fact, double stochasticity prevents from using natural broadcast schemes, in which a given node may transmit its local estimate to *all* neighbors without expecting any immediate feedback.

Remarkably, although generally assumed, double stochasticity of the matrices W_n is in fact **not** mandatory. A couple

*This work is supported by DGA (French Armement Procurement Agency), the Institut Mines-Telecom and by the ANR grant ODISSEE of program ASTRID.

G. Morral, P. Bianchi and G. Fort are with LTCI, Télécom ParisTech & CNRS, 46 rue Barrault, 75634 Paris Cedex 13, France `[firstname].[lastname]@telecom-paristech.fr`

of works (see *e.g.*, [8], [18]) get rid of the column-stochasticity condition, but at the price of assumptions that may not always be satisfied in practice. Other works [**?**], [9] manage to circumvent the use of feedback links by coupling the gradient descent with the so-called push-sum protocol [19]. The latter however introduces an additional communication of weights in the network in order to keep track of some summary of the past transmissions.

The following questions remain unanswered. "What conditions on the sequence $(W_n)_{n\geq 0}$ are needed to ensure that Algorithm (1)-(2) drives all agents to a common critical point of $\sum_i f_i$? What happens if these conditions are not satisfied?".

Contributions.

1) Assuming that $(W_n)_{n\geq 0}$ forms an i.i.d. sequence of stochastic matrices, we prove under some technical hypotheses that Algorithm (1)-(2) leads the agents to a consensus, which is characterized. It is shown that the latter consensus does not necessarily coincide with a critical point of $\sum_i f_i$.
2) We provide sufficient conditions either on the communication protocol *or* on the functions f_i which ensure that limit points are the critical points of $\sum_i f_i$. When such conditions are not satisfied, we also propose a simple modification of the algorithm which allows to recover the sought behavior.
3) We extend our results to a broader setting, assuming that the matrices $(W_n)_{n\geq 0}$ are no longer i.i.d., but are likely to depend on both the current observations and the past estimates. We also investigate a general stochastic approximation framework which goes beyond the model (4) and beyond the only problem of distributed optimization.

The paper is organized as follows. Section II addresses the convergence of a distributed stochastic algorithm. The convergence result is extended to the general setting of distributed stochastic approximation in Section III. The sketch of the proof is given in Section IV. The paper is finally concluded in Section V by numerical illustrations of our results.

Notation: Throughout the paper, the vectors are column vectors. For any vector $x \in \mathbb{R}^\ell$, $|x|$ represents the Euclidean norm of x. I_N is the $N \times N$ identity matrix. $J := \mathbf{1}\mathbf{1}^T/N$ denotes the orthogonal projector onto the linear span of the all-one $N \times 1$ vector $\mathbf{1}$, and $J_\perp := I_N - J$.
The random variables $W_n \in \mathbb{R}^{N\times N}$ and $Y_n := (Y_{n,1}^T, \ldots, Y_{n,N}^T)^T \in \mathbb{R}^{dN}$, $n \geq 1$, are defined on the same measurable space equipped with a probability $\mathbb{P}$; $\mathbb{E}$ denotes the associated expectation. For any $n \geq 1$, define the σ-field $\mathcal{F}_n := \sigma(\theta_0, W_1, \ldots, W_n, Y_1, \ldots, Y_n)$ where θ_0 is the (possibly random) initial point of the algorithm.
Throughout the paper, it is assumed that for any $i \in 1, \ldots, N$, $(\theta_{n,i})_{n\geq 0}$ satisfies the update equations (1)-(2); and we set

$$\theta_n := (\theta_{n,1}^T, \ldots, \theta_{n,N}^T)^T.$$

II. DISTRIBUTED OPTIMIZATION

A. Framework

In this section, we consider the case when $Y_{n,i}$ satisfies (4) with

Assumption 1. 1) *$f_i : \mathbb{R}^d \to \mathbb{R}$ is differentiable and ∇f_i is locally Lipschitz-continuous.*
2) *For any Borel set A of $\mathbb{R}^{dN}$, almost-surely*

$$\mathbb{P}\left[\xi_{n+1} \in A \,|\, \mathcal{F}_n\right] = \nu_{\theta_n}(A)\,,$$

where $(\nu_\theta)_{\theta\in\mathbb{R}^{dN}}$ is a family of probability measures such that
   a) $\int z\, d\nu_\theta(z) = 0$
   b) $\sup_{\theta\in\mathcal{K}} \int |z|^2 d\nu_\theta(z) < \infty$ *for any compact set* $\mathcal{K} \subset \mathbb{R}^{dN}$.

We consider the following assumption on the communication matrix W_n:

Assumption 2. 1) *For any $n \geq 0$, conditionally to $\mathcal{F}_n$, (W_{n+1}, Y_{n+1}) are independent.*
2) *$(W_n)_{n\geq 1}$ is an independent and identically distributed (i.i.d.) sequence of row-stochastic matrices (i.e. $W_n\mathbf{1} = \mathbf{1}$ for any n) with non-negative entries.*
3) *The spectral radius of the matrix $\mathbb{E}[W_1^T J_\perp W_1]$ is strictly lower than* 1.

Assumption 2 implies that at each time n, the communication matrix W_n is random and does not depend on the observations $(Y_1, \ldots, Y_n)$.

The row-stochasticity assumption is a rather mild condition. It claims that $\sum_j w_n(i,j) = 1$ for any i i.e. each node i computes a weighted average of the temporary updates at each node (with possibly some null weights). In many works, it is usually also assumed that W_n is column-stochastic *i.e.* $\sum_i w_n(i,j) = 1$ for any j. Our weaker framework addresses more general gossip protocols, usually less demanding in terms of scheduling and overall network coordination.

Assumption 2-3) is a contraction condition which is required to drive the network to a consensus.

We introduce some assumptions on the step-size sequence $(\gamma_n)_{n\geq 1}$, which are satisfied for example by polynomially decreasing sequences $\gamma_n = \gamma_\star/n^a$ for some $a \in (1/2, 1]$ and $\gamma_\star > 0$.

Assumption 3. *The deterministic step size sequence $(\gamma_n)_{n\geq 1}$ satisfies $\gamma_n > 0$ and:*
1) $\sum_n \gamma_n = +\infty$, $\sum_n \gamma_n^{1+\lambda} < \infty$ *for some* $\lambda \in (0,1)$.
2) $\lim_n \gamma_{n+1}/\gamma_n = 1$ *and* $\sum_n \gamma_n |\gamma_{n+1}/\gamma_n - 1| < \infty$.

Finally, we introduce a stability-like condition.

Assumption 4. *Almost surely, there exists a compact set $\mathcal{K}$ of $\mathbb{R}^{dN}$ such that $\theta_n \in \mathcal{K}$ for any $n \geq 0$.*

Assumption 4 claims that the sequence $(\theta_n)_{n\geq 0}$ remains in a compact set and this compact set may depend on the path. It is implied by the stronger assumption "there exists a compact

set $\mathcal{K}$ of $\mathbb{R}^{dN}$ such that with probability one, $\theta_n \in \mathcal{K}$ for any $n \geq 0$". Checking Assumption 4 is not always an easy task. As the main scope of this paper is the analysis of convergence rather than stability, it is taken for granted: we refer to [8] for sufficient conditions implying stability.

B. Main Result

The statement of our convergence result is prefaced with the following lemma, which shows that the matrix

$$\overline{W} := \mathbb{E}[W_1]$$

admits a unique left Perron eigenvector v; this vector will play a role in the characterization of the limiting points of the gossip algorithm (1)-(2).

Lemma 1. *Under Assumption 2-3), the $\mathbb{R}^N$-valued vector v defined by*

$$v^T := \frac{1}{N}\mathbf{1}^T\overline{W}(I_N - J_\perp\overline{W})^{-1} \tag{5}$$

is the unique non-negative vector satisfying $v^T = v^T\overline{W}$ and $v^T\mathbf{1} = 1$.

Proof: By the Jensen's inequality, for any $x \in \mathbb{R}^N$, $x^T\overline{W}^T J_\perp \overline{W} x \leq x^T\mathbb{E}\left[W_1^T J_\perp W_1\right] x$. Then, by Assumption 2-3), the spectral norm of $J_\perp\overline{W}$ is strictly lower than one. Therefore, $I_N - J_\perp\overline{W}$ is invertible.
The vector v satisfies $v^T\mathbf{1} = 1$ and $v^T\overline{W} = v^T$; to that goal, observe that $(I_N - J_\perp\overline{W})^{-1}\mathbf{1} = \mathbf{1}$. Let us prove that a vector satisfying these two properties is unique; let $w \in \mathbb{R}^N$ satisfying these properties. Then, $w^T = w^T\overline{W} = w^T J_\perp\overline{W} + \mathbf{1}^T\overline{W}/N$ thus implying that $w^T = v^T$.
Since $\overline{W}$ is a stochastic matrix, its spectral radius is one. By [21], there exists a non-negative vector w such that $w^T\overline{W} = w^T$ and $\mathbf{1}^T w > 0$. We can therefore assume without loss of generality that $w^T\mathbf{1} = 1$. The above discussion implies that $w = v$. This concludes the proof. ■

Theorem 1. *Let Assumptions 1, 2, 3 and 4 hold true. Define the function $V : \mathbb{R}^d \to \mathbb{R}$*

$$V(\theta) := \sum_{i=1}^N v_i\, f_i(\theta) \tag{6}$$

where $v = (v_1, \ldots, v_N)$ is the vector defined in Lemma 1. Assume that the set $\mathcal{L} = \{\theta \in \mathbb{R}^d \,|\, \nabla V = 0\}$ of critical points of V is nonempty, bounded, included in some level set $\{\theta : V(\theta) \leq C\}$ and that $V(\mathcal{L})$ has an empty interior. Assume also that the level sets $\{\theta : V(\theta) \leq C\}$ are either empty or compact. The following holds with probability one:

1) *The algorithm converges to a consensus* i.e., $\lim_{n\to\infty} \max_{i,j} |\theta_{n,i} - \theta_{n,j}| = 0$.
2) *The sequence $(\theta_{n,1})_{n\geq 0}$ converges to $\mathcal{L}$ as $n \to \infty$.*

C. Success and Failure of Convergence

Theorem 1 implies that the Algorithm (1)-(2) generally fails to converge towards a critical point of the problem (3). Instead, the algorithm converges to $\mathcal{L}$ which in general is not the set of the critical points of $\theta \mapsto \sum_i f_i(\theta)$. We now discuss some examples of cases when the algorithm does converges to the sought points.

Scenario 1. *All functions f_i are strictly convex and admit a (unique) common minimizer $\theta_\star$.*

This case is for instance investigated by [7] in the framework of statistical estimation in wireless sensor network. In this scenario, we may assume without loss of generality that $f_i(\theta) \geq f_i(\theta_\star) = 0$ for all i (note that the Algorithm (1)-(2) is not modified when f_i is translated). Since $v_i \geq 0$, V is a non-negative strictly convex function such that $V(\theta_\star) = 0$. Therefore, the set of minimizers of V is $\{\theta_\star\}$. On the other hand, since V is convex, $\mathcal{L}$ is the set of minimizers of V. This implies that the set $\mathcal{L}$ *is formed by the minimizers of* $\sum_i f_i$.
The same conclusion holds by relaxing the strict convexity assumption on the functions f_i: if the functions f_i are convex with a common minimizer and $v_i > 0$ for any i, then $\mathcal{L}$ is formed by the minimizers of $\sum_i f_i$. The proof is along the same lines and is omitted.

Scenario 2. *$\overline{W}$ is column-stochastic* i.e., $\mathbf{1}^T\overline{W} = \mathbf{1}^T$.

In this case, v given by Lemma 1 is the vector $\frac{1}{N}\mathbf{1}$. Consequently, $V = \frac{1}{N}\sum_i f_i$. Here again, $\mathcal{L}$ is the set of minimizers of $\sum_i f_i$. An example of random communication protocol satisfying $\mathbf{1}^T\overline{W} = \mathbf{1}^T$ is the following: at time n, a single node i wakes up at random with probability p_i 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, any node k which does not belong to the neighborhood of i (including i itself) sets $\theta_{n,k} = \tilde{\theta}_{n,k}$. Then, given i wakes up, the (k,ℓ)th entry of W_n is given by:

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

W_n is *not* doubly stochastic. However, when nodes wake up according to the uniform distribution ($p_i = \frac{1}{N}$ for all i) it is easily seen that $\mathbf{1}^T\mathbb{E}[W_n] = \mathbf{1}^T$.

Remark 1. *We end up this section with a simple modification of the initial algorithm in the case where $v_i > 0$ for all i. Let us replace the local step (1) of the algorithm by*

$$\tilde{\theta}_{n,i} := \theta_{n-1,i} + \gamma_n\, v_i^{-1} Y_{n,i} \tag{7}$$

where $Y_{n,i}$ is still given by (4). As an immediate Corollary of Theorem 1, the Algorithm (7)-(2) drives the agent to a consensus which coincides with the critical points of $\sum_i f_i$. Note that this modification requires for each node i to have some prior knowledge of the communication protocol through the coefficients v_i (questions related to the practical computation of v_i are however beyond the scope of this paper).

III. A GENERAL ROBBINS-MONRO ALGORITHM

In this section, we consider the general setting described by Algorithm (1)-(2) with weaker conditions on the distribution of the observations Y_n. We also weaken the assumptions on

the conditional distribution of (Y_{n+1}, W_{n+1}) given the past behavior of the algorithm $\mathcal{F}_n$: our general framework includes the case when the communication protocol is adapted at each time n and takes into account the network observations.

For a matrix A, the spectral norm is denoted by $\|A\|$. For any vector $x \in \mathbb{R}^{dN}$ of the form $x = (x_1^T, \ldots, x_N^T)^T$ where $x_i \in \mathbb{R}^d$, we define the vector of $\mathbb{R}^d$

$$\langle x \rangle := \frac{x_1 + \cdots + x_N}{N} = \frac{1}{N}(\mathbf{1}^T \otimes I_d)x \,, \tag{8}$$

where $\otimes$ denotes the Kronecker product. Recall the standard formula

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD) \,. \tag{9}$$

We extend the notation to matrices $X \in \mathbb{R}^{dN \times k}$ as $\langle X \rangle := \frac{1}{N}(\mathbf{1}^T \otimes I_d)X \in \mathbb{R}^{d \times k}$. Set $\mathcal{J} = J \otimes I_d$ and $\mathcal{J}_\perp = J_\perp \otimes I_d$. Matrix $\mathcal{J}$ represents the orthogonal projector onto the *consensus space* defined as the set of vectors $x \in \mathbb{R}^{dN}$ whose d-dimensional blocs $x_1, \ldots, x_N$ are all equal. As a consequence of (9), $\mathcal{J}x = \mathbf{1} \otimes \langle x \rangle$. Following this notation, we define the gossip matrix as $\mathcal{W}_n = W_n \otimes I_d$. The Algorithm (1)-(2) under study can be written as the following matrix form:

$$\theta_n = \mathcal{W}_n \left(\theta_{n-1} + \gamma_n Y_n\right) \,. \tag{10}$$

The following assumption describes the distribution of (Y_{n+1}, W_{n+1}) given the past $\mathcal{F}_n$. We denote by M_1 the set of $N \times N$ non-negative row-stochastic matrices and we endow M_1 with its Borel σ-field.

Assumption 5. 1) *There exists a collection of distributions $(\mu_\theta)_{\theta \in \mathbb{R}^{dN}}$ on $\mathbb{R}^{dN} \times \mathsf{M}_1$ such that for any Borel set A:*

$$\mathbb{P}\left[(Y_{n+1}, W_{n+1}) \in A \,|\, \mathcal{F}_n\right] = \mu_{\theta_n}(A) \qquad \textit{almost-surely.}$$

In addition, the application $\theta \mapsto \mu_\theta(A)$ defined on $\mathbb{R}^{dN}$ is measurable for any A in the Borel σ-field of $\mathbb{R}^{dN} \times \mathsf{M}_1$.

2) $\sup_{\theta \in \mathcal{K}} \int |y|^2 d\mu_\theta(y, w) < \infty$ *for any compact set $\mathcal{K} \subset \mathbb{R}^{dN}$.*

Assumption 5-1) means that the joint distribution of the r.v.'s Y_{n+1} and W_{n+1} depends on the past $\mathcal{F}_n$ only through the last value θ_n of the vector of estimates. It also implies that W_n is almost-surely (a.s.) non-negative and row-stochastic. Since the variables (Y_{n+1}, W_{n+1}) are not necessarily conditionally independent and $(W_n)_{n \geq 0}$ are no more i.i.d., the contraction condition on $J_\perp W_1$ is replaced with the following condition:

Assumption 6. *For any compact set $\mathcal{K} \subset \mathbb{R}^{dN}$, there exists $\rho_\mathcal{K} \in (0,1)$ such that for all $\theta \in \mathcal{K}$, ϕ in $\mathbb{R}^{dN}$ and $A \in \mathbb{R}^{dN \times dN}$,*

$$\int (\phi + Ay)^T \mathcal{W}^T \mathcal{J}_\perp \mathcal{W}(\phi + Ay) \; d\mu_\theta(y, w) \leq \rho_\mathcal{K} \int |\phi + Ay|^2 \, d\mu_\theta(y, w) \,,$$

where $\mathcal{W} := (w \otimes I_d)$.

Assumption 6 implies that

$$\sup_{\theta \in \mathcal{K}} \sup_{x, |x|=1} \left| \int \mathcal{J}_\perp \mathcal{W} x \; d\mu_\theta(y, w) \right|^2 \leq \sup_{\theta \in \mathcal{K}} \left| \int \mathcal{W}^T \mathcal{J}_\perp \mathcal{W} \; d\mu_\theta(y, w) \right| \leq \rho_\mathcal{K} \,, \tag{11}$$

Regularity conditions on the conditional distribution of the input variables $(Y_n, W_n)_n$ are also required.

Assumption 7. *For any compact set $\mathcal{K} \subset \mathbb{R}^{dN}$, there exists a constant $C_\mathcal{K} > 0$ such that for any $\theta, \theta' \in \mathcal{K}$,*

$$\left| \int w \, d\mu_\theta(y, w) - \int w \, d\mu_{\theta'}(y, w) \right| \leq C_\mathcal{K} |\theta - \theta'| \,, \tag{12}$$

$$\left| \int \langle \mathcal{W} y \rangle \, d\mu_\theta(y, w) - \int \langle \mathcal{W} y \rangle \, d\mu_{\mathcal{J}\theta}(y, w) \right| \leq C_\mathcal{K} |\mathcal{J}_\perp \theta| \,, \tag{13}$$

$$\left| \int \mathcal{J}_\perp \mathcal{W} y \, d\mu_\theta(y, w) - \int \mathcal{J}_\perp \mathcal{W} y \, d\mu_{\theta'}(y, w) \right| \leq C_\mathcal{K} |\theta - \theta'| \,, \tag{14}$$

where $\mathcal{W} := w \otimes I_d$.

Let us introduce the following quantities for any $\theta \in \mathbb{R}^{dN}$:

$$\Omega_\theta := \int \mathcal{J}_\perp \mathcal{W} \, d\mu_\theta(y, w) \tag{15}$$

$$\upsilon_\theta := \int \mathcal{J}_\perp \, \mathcal{W} \, y \, d\mu_\theta(y, w) \tag{16}$$

$$m_\theta := (I_{dN} - \Omega_\theta)^{-1} \upsilon_\theta \,, \tag{17}$$

where $\mathcal{W} := w \otimes I_d$. Under Assumption 5-2), it is not difficult to show that for any compact $\mathcal{K} \subset \mathbb{R}^{dN}$, $\sup_{\theta \in \mathcal{K}} \|\Omega_\theta\| \leq \sqrt{\rho_\mathcal{K}}$, which implies that m_θ is well defined. Finally, define the so-called *mean-field* function $h : \mathbb{R}^d \to \mathbb{R}^d$ by

$$h(\vartheta) = \int \langle \mathcal{W}(y + m_{\mathbf{1} \otimes \vartheta}) \rangle d\mu_{\mathbf{1} \otimes \vartheta}(y, w) \,, \tag{18}$$

where $\mathcal{W} := w \otimes I_d$. We finally assume that there exists a Lyapunov function V for the mean field function h.

Assumption 8. 1) *$h : \mathbb{R}^d \to \mathbb{R}^d$ is continuous.*

2) *there exists a continuously differentiable function $V : \mathbb{R}^d \to \mathbb{R}^+$ such that*
   a) *there exists $M > 0$ such that $\mathcal{L} := \{\vartheta \in \mathbb{R}^d : \nabla V^T(\vartheta) h(\vartheta) = 0\} \subset \{V \leq M\}$. In addition, $V(\mathcal{L})$ has an empty interior.*
   b) *there exists $M' > M$ such that $\{V \leq M'\}$ is a compact subset of $\mathbb{R}^d$.*
   c) *for any $\vartheta \in \mathbb{R}^d \setminus \mathcal{L}$, $\nabla V^T(\vartheta) h(\vartheta) < 0$.*

Assumptions 5 and 6 imply that $\vartheta \mapsto m_{\mathbf{1} \otimes \vartheta}$ is continuous on $\mathbb{R}^d$. Therefore, a sufficient condition for Assumption 8-1) is to strengthen the conditions (13-14) as follows

$$\left| \int \mathcal{W} y \, d\mu_\theta(y, w) - \int \mathcal{W} y \, d\mu_{\theta'}(y, w) \right| \leq C_\mathcal{K} |\theta - \theta'| \,.$$

Observe that when V is a continuous coercive function *i.e.*, a continuous function such that $\lim_{|\vartheta| \to \infty} V(\vartheta) = \infty$, then the level sets $\{V \leq M\}$ are compact subsets of $\mathbb{R}^d$.

Theorem 2. *Let Assumptions 3, 4, 5, 6, 7 and 8 hold true for the algorithm defined by* (10)*. The following properties hold with probability one:*

1) *The algorithm converges to a consensus* i.e., $\lim_{n\to\infty} \mathcal{J}_\perp \theta_n = 0$*;*
2) *The sequence* $(\theta_{n,1})_{n\geq 0}$ *converges to a connected component of* $\mathcal{L}$*, where* $\mathcal{L}$ *is defined in Assumption 8-2).*

Theorem 1 can be obtained as a special case of Theorem 2. Indeed, the assumptions of Theorem 1 imply the assumptions 5 to 8 as we now prove.
For any $\theta = (\theta_1, \ldots, \theta_N) \in \mathbb{R}^{dN}$ where $\theta_i \in \mathbb{R}^d$, define the $\mathbb{R}^{dN}$-valued function g by

$$g(\theta) := (-\nabla f_1(\theta_1)^T, \ldots, -\nabla f_N(\theta_N)^T)^T .$$

Under Assumption 2-1) and Assumption 2-2), for any Borel set $A \times B$ of $\mathbb{R}^{dN} \times \mathsf{M}_1$

$$\mathbb{P}[(Y_{n+1}, W_{n+1}) \in A\times B|\mathcal{F}_n] = \mathbb{P}[Y_{n+1} \in A|\mathcal{F}_n]\mathbb{P}[W_{n+1} \in B] .$$

In addition, by Assumption 1 and Eq. (4)

$$\mathbb{P}[Y_{n+1} \in A|\mathcal{F}_n] = \int \mathbb{I}_A \left(g(\theta_n) + z\right) \; d\nu_{\theta_n}(z) ,$$

where $\mathbb{I}_A$ denotes the indicator function of a set A. This above discussion provides the expression of μ_θ in Assumption 5. In addition, under Assumption 1-2), for any compact set $\mathcal{K}$ of $\mathbb{R}^{dN}$,

$$\sup_{\theta\in\mathcal{K}} \int |y|^2 d\mu_\theta(y, w) = \sup_{\theta\in\mathcal{K}} \left(|g(\theta)|^2 + \int |z|^2 d\nu_\theta(z) \right) < \infty$$

which proves Assumption 5-2). The above expression of μ_θ implies that

$$\int (\phi + Ay)^T \mathcal{W}^T \mathcal{J}_\perp \mathcal{W}(\phi + Ay) \; d\mu_\theta(y, w)$$
$$= \int (\phi+A(g(\theta)+z))^T \mathbb{E}\left[\mathcal{W}^T \mathcal{J}_\perp \mathcal{W}\right] (\phi+A(g(\theta)+z)) \, d\nu_\theta(z) .$$

Therefore, Assumption 6 easily follows from Assumption 2-3). The regularity conditions of Assumption 7 are satisfied by Assumption 1 as the left hand side of (13) is zero and (12) and (14) are true as long as $(\nabla f_i)_i$ are local Lipschitz-continuous. Again, the expression of μ_θ implies that

$$\Omega_\theta = \mathcal{J}_\perp \, \mathbb{E}\left[\mathcal{W}_1\right] ,$$
$$\upsilon_\theta = \mathcal{J}_\perp \, \mathbb{E}\left[\mathcal{W}_1\right] \, g(\theta) .$$

Therefore, the mean field vector h defined by (18) gets into $h(\vartheta) = \langle \mathbb{E}[\mathcal{W}_1] \, \mathcal{A} \, g\left(\mathbf{1} \otimes \vartheta\right)\rangle$ where

$$\mathcal{A} := \left(I_{dN} + (I_{dN} - \mathcal{J}_\perp \mathbb{E}\left[\mathcal{W}_1\right])^{-1} \mathcal{J}_\perp \mathbb{E}\left[\mathcal{W}_1\right]\right) .$$

Using the Woodbury matrix identity (see [21]), we have

$$h(\vartheta) = (v^T \otimes I_d) \; g(\mathbf{1} \otimes \vartheta) = -\sum_{i=1}^N v_i \nabla f_i(\vartheta)$$

where $v = (v_1, \ldots, v_N)$ is given by Lemma 1. Set $\bar{V} := \exp(V)$ where V is defined by (6). Upon noting that $\nabla \bar{V} = -h \, \bar{V}$, it is easily seen that under the assumptions of Theorem 1, Assumption 8 holds.

IV. SKETCH OF THE PROOF OF THEOREM 2

We provide here the sketch of the proof of Theorem 2. The detailed proof will be provided in an extended version of this paper.

By using (8), we write

$$\theta_n = \mathbf{1} \otimes \langle \theta_n \rangle + \mathcal{J}_\perp \theta_n .$$

We define the normalized disagreement vector $(\phi_n)_{n\geq 0}$ by

$$\phi_n = \gamma_{n+1}^{-1} \, \mathcal{J}_\perp \theta_n \quad \text{where} \quad \alpha_n = \gamma_n / \gamma_{n+1} . \tag{19}$$

The following lemma establishes the dynamics of the consensus sequence $(\langle \theta_n \rangle)_{n\geq 0}$ and of the normalized disagreement sequence $(\phi_n)_{n\geq 0}$.

Lemma 2. *Let* $(\theta_n)_{n\geq 0}$ *be the sequence given by (10). Assume that* $(W_n)_{n\geq 0}$ *are row-stochastic matrices. It holds*

$$\langle \theta_n \rangle = \langle \theta_{n-1} \rangle + \gamma_n \langle \mathcal{W}_n (Y_n + \phi_{n-1}) \rangle \; , \tag{20}$$
$$\phi_n = \alpha_n \, \mathcal{J}_\perp \mathcal{W}_n (\phi_{n-1} + Y_n) . \tag{21}$$

The next step is to prove that the square norm of the normalized disagreement sequence is uniformly bounded, conditionally on the event that the past values of the sequence $(\theta_j)_{j\geq 0}$ remain in a compact set. A main ingredient for the proof of this lemma is the contraction property Assumption 6.

Lemma 3. *Let Assumptions 5 and 6 hold true. For any compact set* $\mathcal{K} \subset \mathbb{R}^{dN}$*,*

$$\sup_n \mathbb{E}\left[|\phi_n|^2 \mathbb{I}_{\bigcap_{j\leq n-1}\{\theta_j \in \mathcal{K}\}} \right] < \infty .$$

Since $\sum_n \gamma_n^{1-\lambda} < \infty$ for some $\lambda \in (0, 1)$ under Assumption 3, this lemma implies the convergence result on the disagreement vector:

Proposition 1 (Agreement)**.** *Let Assumptions 3-1), 4, 5 and 6 hold true. Then almost-surely,* $\lim_{n\to\infty} \mathcal{J}_\perp \theta_n = 0$.

The second step is to address the long-time behavior of the average estimate $\langle \theta_n \rangle$. To that goal, we write the update rule (20) as a stochastic approximation algorithm. We have

$$\langle \theta_n \rangle = \langle \theta_{n-1} \rangle + \gamma_n \mathbb{E}\left[\eta_n|\mathcal{F}_{n-1}\right] + \gamma_n \left(\eta_n - \mathbb{E}\left[\eta_n|\mathcal{F}_{n-1}\right]\right) ,$$

where $\eta_n := \langle \mathcal{W}_n (Y_n + \phi_{n-1}) \rangle$. Under the stated assumptions on the conditional distribution of (Y_{n+1}, W_{n+1}) given the past $\mathcal{F}_n$, $\mathbb{E}\left[\eta_n|\mathcal{F}_{n-1}\right] = \int \langle \mathcal{W}\left(y + \phi_{n-1}\right) \rangle \, d\mu_{\theta_{n-1}}(y, w)$ where $\mathcal{W} := w \otimes I_d$. By Proposition 1, θ_{n-1} and $\mathbf{1} \otimes \langle \theta_{n-1} \rangle$ are close when n is large so

$$\mathbb{E}\left[\eta_n|\mathcal{F}_{n-1}\right] = \int \langle \mathcal{W}\left(y + \phi_{n-1}\right) \rangle \, d\mu_{\mathbf{1}\otimes\langle\theta_{n-1}\rangle}(y, w) + \Xi_{n-1}^{(1)} \tag{22}$$

where $(\Xi_n^{(1)})_n$ is a remainder term in an appropriate sense. In addition, (21) shows that $(\phi_n)_{n\geq 1}$ is a controlled Markov chain:

$$\mathbb{P}(\phi_n \in A|\mathcal{F}_{n-1}) = P_{\theta_{n-1},\alpha_n}(\phi_{n-1}, A)$$

where $P_{\theta,\alpha}(x, A) := \int \mathbb{I}_A \left(\alpha \mathcal{J}_\perp \mathcal{W}(x + y)\right) \, d\mu_\theta(y, w)$. We show that under our assumptions, the transition kernel $P_{\theta,\alpha}$

possess a unique invariant distribution $\pi_{\theta,\alpha}$. When n is large, $\alpha_n \sim 1$ and $\lim_n |\theta_n - \mathbf{1} \otimes \langle \theta_n \rangle| = 0$ (almost-surely) so the rough intuition is that in (22), ϕ_{n-1} can be replaced with the expectation of $\pi_{\mathbf{1}\otimes\langle\theta_{n-1}\rangle,1}$; this expectation is $m_{\mathbf{1}\otimes\langle\theta_{n-1}\rangle}$ given by (17). Combining these successive approximations yield $\mathbb{E}[\eta_n|\mathcal{F}_{n-1}] = h(\langle\theta_{n-1}\rangle) + \Xi_{n-1}^{(1)} + \Xi_{n-1}^{(2)}$ where h is given by (18) and $(\Xi_n^{(2)})_n$ is small in some appropriate sense. We then establish the convergence results by verifying the sufficient conditions of [22, Theorem 2.2 and 2.3] for the convergence of stochastic approximation algorithms.

V. NUMERICAL RESULTS

Consider a network of $N = 5$ agents and for any $i = 1, \ldots, 5$, we define $f_i : \mathbb{R} \to \mathbb{R}$ by $f_i(\theta) = \frac{1}{2}(\theta - \lambda_i)^2$ where $\lambda^T = (1, 4, -2, 2, 0)$. The minimizer of $\sum_i f_i$ is $\theta_f = 1$. Consider the graph with vertices $\{1, \ldots, N\}$ and edges $\{1, 2\}$, $\{2, 3\}$, $\{3, 4\}$, $\{4, 5\}$, $\{3, 5\}$. We choose $\theta_{0,i} = 0$ for all agent i, $\gamma_n = 1/n$. Figure 1 represents a realization of the sequence $(\theta_{n,1})_{n\geq 0}$ as a function of n in 3 different scenarios. The plain line curve with square markers corresponds to the standard algorithm (10) when $Y_{n,i}$ is defined as in (4): $(\xi_{n,i})_{n,i}$ is a i.i.d. sequence with Gaussian distribution $\mathcal{N}(0, 1)$ and W_n is assumed fixed and deterministic ($W_n = W_1$); we select W_1 in such a way that each agent computes the average of the temporary estimates in its neighborhood. This is equivalent to set $W_1 = (D+I)^{-1}(A+I)$, where A is the adjacency matrix of the graph and where D is the diagonal matrix containing the degrees. Note that $\mathbf{1}^T W_1 = \frac{1}{12}(10, 13, 15, 11, 11) \neq \mathbf{1}^T$. Computing the left Perron eigenvector (see (5)) yields $\frac{1}{5}v^T = (\frac{2}{3}, 1, \frac{4}{3}, 1, 1)$. The minimizer of $V = \sum_i v_i f_i$ is $\theta_V = 0.8$. Figure 1 shows that the sequence $(\theta_{n,1})_n$ converges to θ_V. The plain line curve with circle markers on Figure 1 represents the same algorithm when $Y_{n,i}$ is replaced by $v_i^{-1}(-\nabla f_i(\theta_{n-1,i}) + \xi_{n,i})$ as in Remark 1. We refer to this algorithm as *weighted*. As predicted by the Theorem 1-1), the algorithm now converges to the sought value $\theta_f = 1$. Finally, the plain line curve represents the case where the broadcast gossip protocol depicted in Scenario 2 is used with uniform node selection $p_i = \frac{1}{N}$ and with $\beta = \frac{1}{2}$. The algorithm converges to the sought value θ_f as discussed in Section II-C.

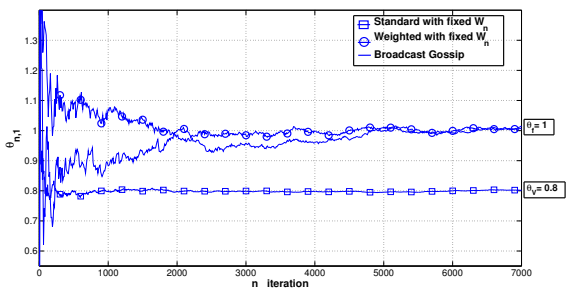

Figure 1: Trajectory of $\theta_{n,1}$ as a function of n.

Figure 2 represents the norm of the scaled disagreement vector as a function of n for the same algorithms. As expected from Theorem 1-2), consensus is asymptotically achieved.

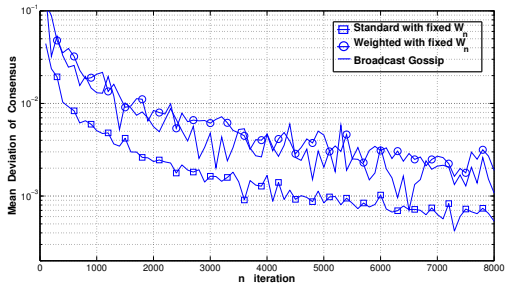

Figure 2: $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\theta_{n,i} - \langle\theta_n\rangle)^2}$ as a function of n

REFERENCES

[1] M. G. Rabbat and R. D. Nowak, "Quantized Incremental Algorithms for Distributed Optimization," *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 4, pp. 798–808, 2005.

[2] S. Ram, A. Nedic, and V. Veeravalli, "Incremental Stochastic Subgradient Algorithms for Convex Optimization," *SIAM Journal on Optimization*, vol. 20, no. 2, pp. 691–717, 2009.

[3] J. Tsitsiklis, "Problems in Decentralized Decision Making and Computation," Ph.D. dissertation, Massachusetts Institute of Technology, 1984.

[4] 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.

[5] S. Kar and J. 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] S. Stankovic and, M. Stankovic, and D. Stipanovic, "Decentralized Parameter Estimation by Consensus Based Stochastic Approximation," *IEEE Transactions on Automatic Control*, vol. 56, no. 3, pp. 531–543, march 2011.

[7] J. Chen and A. Sayed, "Diffusion adaptation strategies for distributed optimization and learning over networks," *IEEE Trans. Signal Processing*, vol. 60, no. 8, pp. 4289–4305, May 2012.

[8] P. Bianchi, G. Fort, and W. Hachem, "Performance of a Distributed Stochastic Approximation Algorithm," *IEEE Trans. on Information Theory*, vol. 59, no. 11, pp. 7405–7418, 2012.

[9] A. Nedic and A. Olshevsky, "Distributed optimization over time-varying directed graphs," in *IEEE conf. on Decision and Control*, Florence, Italy, 2013.

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

[11] J. Chen, C. Richard, and A. Sayed, "Multitask diffusion adaptation over networks," *Signal Processing, IEEE Transactions on*, 2014 (submitted).

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

[13] A. Abboud, R. Couillet, M. Debbah, and H. Siguerdidjane, "Asynchronous alternating direction method of multipliers applied to the direct-current optimal power flow problem," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, Italy, 2014.

[14] 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.

[15] S. 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. [Online]. Available: http://dx.doi.org/10.1007/s10957-010-9737-7

[16] 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.

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

[18] A. Nedic, "Asynchronous Broadcast-Based Convex Optimization Over a Network," *IEEE Transactions on Automatic Control*, vol. 56, no. 6, pp. 1337–1351, june 2011.

[19] D. Kempe, A. Dobra, and J. Gehrke, "Gossip-Based Computation of Aggregate Information." IEEE Computer Society, 2003.

[20] F. Bénézit, V. Blondel, P. Thiran, J. Tsitsiklis, and M. Vetterli, "Weighted gossip: Distributed averaging using non-doubly stochastic matrices," in *Information Theory Proceedings (ISIT), IEEE International Symposium on*. IEEE, 2010, pp. 1753–1757.

[21] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1990.

[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.