Distributed Stochastic Approximation: The Price of Non-double Stochasticity

Gemma Morral* Pascal Bianchi* Gersende Fort* Jérémie Jakubowicz†
*CNRS/LTCI Telecom ParisTech, Paris, France. E-mail: {firstname.lastname}@telecom-paristech.fr
†CNRS/SAMOVAR Telecom Sud-Paris, Evry, France. E-mail: {firstname.lastname}@telecom-sudparis.fr

***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 at each agent and a gossip step which drives the network to a consensus. The gossip step uses row-stochastic matrices to weight network exchanges. Gossip-matrices are often also assumed column-stochastic in the literature. Unfortunately, column-stochasticity implies significant restrictions on the communication protocol and prevents from using simple broadcast protocols. Under the assumption of decreasing step sizes, it is proved that the network is driven to a consensus at overwhelming speed and that the average estimate converges to the sought consensus. When the gossip matrices are doubly stochastic, a central limit theorem is established and it is proved that the performance of the algorithm is identical to that of a centralized algorithm. When the gossip matrices are non doubly stochastic, an excess variance term is added to the limiting distribution. In that case, a performance gap w.r.t. the centralized algorithm exists and is characterized.**

I. Introduction

Stochastic approximation (SA) 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 a sequence of parameters, Y_n is a sequence of random variables, 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 gradient.

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

The literature contains at least two different cooperation approaches for solving the distributed optimization problem. In the so-called *incremental* approach (see for instance [3], [4]): a message containing an estimate of the desired minimizer iteratively travels all over the network. At any instant, the agent which is in possession of the message updates its own estimate and adds its own contribution, based on its local observation. Incremental algorithms generally require the message to go through a Hamiltonian cycle in the network. Finding such a path is known to be a NP complete problem and is not particularly suitable to distributed computations. Relaxations of the Hamiltonian cycle requirement have been proposed: for instance, [4] only requires that an agent communicates with another agent randomly selected in the network (not necessarily in its neighborhood) according to the uniform distribution. However, substantial routing is still needed. In [5], distributed optimization is tackled using a different approach, assuming that agents perfectly observe their utility functions and know also the utility functions of their neighbors.

This paper focuses on another cooperation approach based on *average consensus* techniques. In this context, each agent maintains its own estimate. Agents separately run local gradient algorithms and simultaneously communicate in order to eventually reach an agreement over the whole network on the value of the minimizer. Communicating agents combine their local estimates in a linear fashion: a receiver computes a weighted average between its own estimate and the ones which have been transmitted by its neighbors. Such combining techniques are often referred to as *gossip* methods.

The idea beyond the algorithm of interest in this paper is not new. Its roots can be found in [6], [7] where a network of processors seeks to optimize some objective function *known* by all agents (possibly up to some additive noise). More recently, numerous works extended this kind of algorithm to more involved multi-agent scenarios, see [8]–[10] as a non exhaustive list. Multi-agent systems are indeed more difficult to deal with, because individual agents ignore the global objective function to be minimized. [11] addresses the problem of *unconstrained* optimization, assuming *convex* but non necessarily differentiable utility functions. Convergence to a global minimizer is established assuming that utility functions have bounded (sub)gradients. Let us also mention [10] which focuses on the case of *quadratic* objective functions. Unconstrained optimization is also investigated in [12] assuming differentiable but non necessarily convex utility functions and relaxing boundedness conditions on the gradients. Convergence to a critical point of the objective function is proved and the asymptotic performance is evaluated under the form of a central limit theorem. In [8], the problem of *constrained* distributed optimization is addressed. Convergence to an optimal consensus is proved

The work of G. Morral is supported by DGA (French Armement Procurement Agency).

when each utility function f_i is assumed convex and perfectly known by agent i. These results are extended in [13] to the stochastic descent case *i.e.*, when the observation of utility functions is perturbed by a random noise.

In each of these works, the gossip communication scheme can be represented by a sequence of matrices $(W_n)_{n\geq 1}$ of size $N\times N$, where the (i,j)th component of W_n is the weight given by agent i to the message received from j at time n, and is equal to zero in case agent i receives no message from j. In most works (see for instance [8], [11]–[13]), matrices W_n 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 [14], each one-way transmission from an agent i to another agent j requires at the same time a feedback link from j to i. Double stochasticity prevents from using natural broadcast schemes, in which a given agent may transmit its local estimate to *all* its neighbors without expecting any immediate feedback [15]. Very recently, [12], [16], [17] get rid the column stochasticity condition and prove that convergence to the sought consensus can indeed be achieved using rather simple communication protocols of broadcast nature.

II. Distributed Stochastic Approximation

A. The Algorithm

In this paper, we consider a network composed by N nodes (sensors, robots, computing units, ...). Node i generates a stochastic process $(\theta_{n,i})_{n\geq 1}$, assumed to be real valued for simplicity: the vector-case will be addressed in an extended version of this paper. At time n, the algorithm under study both involves a local step and a gossip step:

`[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 $\mathbb{R}$-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:

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

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 $W_n\mathbf{1} = \mathbf{1}$.

We refer to (1)-(2) as the distributed stochastic approximation algorithm (DSAA). Define the random vectors θ_n and Y_n as $\theta_n := (\theta_{n,1},\ldots,\theta_{n,N})^T$ and $Y_n = (Y_{n,1},\ldots,Y_{n,N})^T$. The DSAA reduces to:

$$\theta_n = W_n\left(\theta_{n-1} + \gamma_n Y_n\right)\ . \tag{3}$$

B. Observation and Network Models

The random process $(Z_n)_{n\geq 1} := ((Y_n, W_n))_{n\geq 1}$ is defined on a measurable space equipped with a probability $\mathbb{P}$; $\mathbb{E}$ denotes the associated expectation. For any $n\geq 1$, we introduce the σ-field $\mathcal{F}_n = \sigma(\boldsymbol{\theta}_0, Z_{1:n})$.

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

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

Moreover, Y_{n+1} and W_{n+1} are independent conditionally to $\mathcal{F}_n$.

The following function $h:\mathbb{R}\to\mathbb{R}^N$ will be revealed crucial in our analysis:

$$h(\theta) := \int y\,\mu_{\theta\mathbf{1}}\,(dy)\ .$$

The ith component of $h(\theta)$ represents the expectation of the ith agent's observation conditionally to the event that all the agents have the same estimate θ. In particular, we shall see that under some technical conditions, the DSAA drives all agents estimates to a root $\theta^\star$ of the average function $\overline{h}:\mathbb{R}\to\mathbb{R}$

$$\overline{h}(\theta) := \mathbf{1}^T h(\theta)/N\ .$$

We shall refer to $\overline{h}$ as the *mean field* of the DSAA. We define $J := \mathbf{1}\mathbf{1}^T/N$ as the orthogonal projector onto the linear span of $\mathbf{1}$ and $J^\perp := I_N - J$ where I_N is the identity matrix of size N. We set:

$$W_n^\perp := J^\perp W_n\ .$$

We denote by $\rho(M)$ the spectral radius of any square matrix M.

Assumption 2. *The following conditions hold:*

a) $(W_n)_{n\geq 1}$ is an i.i.d. sequence of row-stochastic matrix-valued random variables with non-negative components:

$$\forall n,\ W_n\mathbf{1} = \mathbf{1}\ .$$

b) Matrix $\mathbb{E}(W_1)$ is column-stochastic:

$$\mathbb{E}(W_1)^T\mathbf{1} = \mathbf{1}\ .$$

c)

$$\rho\left(\mathbb{E}\left(W_1^{\perp,T}W_1^\perp\right)\right) < 1\ . \tag{4}$$

The assumption that W_n is row-stochastic for any n is a rather mild condition. It implies that each agent i in (2) computes a weighted average in the sense that

$$\sum_j w_n(i,j) = 1\ .$$

In previous works, it is usually also assumed that $\sum_i w_n(i,j) = 1$ for any j *i.e.*, matrix W_n is column-stochastic. In this paper, we do not use this assumption. As

a consequence, we are able to use more general gossip protocols that are usually less demanding in terms of scheduling and overall network coordination. This point will be further discussed in the next paragraph. Here, we only require that W_n is column-stochastic *in average*. Finally, Assumption 2c) can be seen as a contraction condition which is needed in order to drive the network to a consensus.

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

C. Illustration: Some Examples of Gossip Schemes

Before proceeding with the convergence analysis of algorithm (3), it is worth discussing how the inter-agent communication scheme affects matrices W_n. We describe two standard gossip schemes so called *pairwise* and *broadcast* schemes. The reader can refer to [18] 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 [14] on average consensus. 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 exchange information: $\theta_{n,k} = \tilde{\theta}_{n,k}$. 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 $\mathbb{R}^N$. In particular, matrices $(W_n)_{n\geq 1}$ are i.i.d. and doubly stochastic:

$$W_n \mathbf{1} = \mathbf{1} \ , \qquad W_n^T \mathbf{1} = \mathbf{1} \ .$$

It is shown in [14] that (4) holds if and only if the weighted graph $(\mathsf{E}, \mathsf{V}, \mathsf{W})$ is connected, where the edge $\{i, j\}$ is weighted by the probability that the nodes i, j communicate at time n.

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, any node k which does 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 **2**. Here again, it can be shown that the spectral norm ρ of $\mathbb{E}(W_1^{\perp,T} W_1^{\perp})$ is in $[0, 1)$ if and only if (E, V) is a connected graph (see [15]).

III. Almost-sure convergence of DSAA

For the sake of completeness, this section recalls convergence results previously stated in [12]. For any vector $x \in \mathbb{R}^N$, set $\overline{x} := N^{-1} \sum_{i=1}^N x_i$ and $x^\perp := J^\perp x$. Any vector x can be decomposed as the direct sum $\overline{x}\mathbf{1} + x^\perp$ where the first and second terms are respectively the orthogonal projections of x onto the linear span of $\mathbf{1}$ and the orthogonal hyperplane.

We need further assumptions on the mean-field $\bar{h}$.

Assumption 4. *There exists* $V : \mathbb{R} \to \mathbb{R}^+$ *such that:*

a) V is continuously differentiable, and its first derivative is denoted by V'.
b) For any $\theta \in \mathbb{R}$, $V'(\theta)\overline{h}(\theta) \leq 0$.
c) For any $M > 0$, the level set $\{\theta \in \mathbb{R} : V(\theta) \leq M\}$ is compact.
d) The set $\mathcal{L} := \{\theta \in \mathbb{R} : V'(\theta)\,\overline{h}(\theta) = 0\}$ is such that $V(\mathcal{L})$ has an empty interior.

This assumption says that the dynamical system $\dot{\theta} = \overline{h}(\theta)$ is dissipative: if we think of $V(\theta)$ as an energy associated to θ, then, $V(\theta)$ is decreasing along a trajectory (assumption 4.2). Moreover $\lim_{|\theta|\to\infty} V(\theta) = \infty$ (assumption 4.3) says that it is not possible to keep finite energy while going to infinity. Such a function V is called a *Lyapunov function*.

Assumption 5. *a) There exists $C > 0$ such that for any $\theta \in \mathbb{R}^N$:*

$$\left| \int \bar{y}\mu_\theta(dy) - \int \bar{y}\mu_{\bar{\theta}\mathbf{1}}(dy) \right| \leq C|\theta^\perp|$$

b) $\sup_n \mathbb{E}\left[|Y_n|^2\right] < \infty$.

The first part of the assumption claims that for θ close to consensus (that is $\theta^\perp$ is close to zero), the conditional distribution of Y given the past μ_{θ_n} behaves as $\mu_{\bar{\theta}_n\mathbf{1}}$. The second part is a stability-like condition. Of course, checking Assumption 5b) is not always an easy task. Here, as the paper focuses on convergence rates rather than stability, this Assumption is taken for granted: we refer to [12] for sufficient conditions implying Assumption 5b).

Theorem 1 ([12])**.** *Assume Assumptions 1 to 5. Then* $(\theta_n)_{n\geq 0}$ *converges almost-surely to the set*

$$\{\theta\mathbf{1} \ : \ \theta \in \mathcal{L}\} \ .$$

Theorem 1 implies that under the stated assumptions, an agreement is achieved: the estimates $\theta_{n,i}$ eventually agree for all i in the sense that $\theta_{n,i} - \overline{\theta}_n$ tends to zero, where $\overline{\theta}_n$ represents the average estimate at time n w.r.t. all agents. Moreover $(\overline{\theta}_n)_{n\geq 0}$ converges to the set $\mathcal{L}$. In case $\mathcal{L}$ is reduced to a singleton $\{\theta^\star\}$, Theorem 1 can be restated as: almost-surely,

$$\forall i, \quad \lim_n \theta_{n,i} = \theta^\star \ . \tag{5}$$

IV. Convergence rates

In order to lighten the presentation and avoid tedious conditioning, we assume from now on that $\mathcal{L} = \{\theta^\star\}$ for

some $\theta^\star \in \mathbb{R}$. A generalization will be provided in an extended version of this paper.

In order to analyze the asymptotic behavior of θ_n, it turns out convenient to separately study $\overline{\theta}_n$ and $\theta_n^\perp$. Here, $\overline{\theta}_n$ represents the average estimate of the network and $\theta_n^\perp$ is the so-called *disagreement vector* whose ith component coincides with $\theta_{n,i} - \overline{\theta}_n$.

Assumption 6. *a) There exist $\delta > 0$, $\tau > 0$ such that:*

$$\sup_{|\theta - \theta^\star \mathbf{1}| \leq \delta} \int \mu_\theta(dy) \ |y|^{2+\tau} < \infty \, .$$

b) For any bounded continuous f, the function defined on $\mathbb{R}^N$ by $\theta \mapsto \int f d\mu_\theta$ is continuous at $\theta^\star \mathbf{1}$.

c) The function $Q : \mathbb{R}^N \to \mathbb{R}$ defined by:

$$Q(\theta) := \int y \cdot y^T \mu_\theta(dy)$$

is continuous at $\theta^\star \mathbf{1}$.

A. Normalized Disagreement Vector

We first analyze the normalized disagreement vector $\theta_n^\perp / \gamma_{n+1}$. We set $h^\star := h(\theta^\star)$ and $Q^{\star,\perp} := J^\perp Q(\theta^\star \mathbf{1}) J^\perp$. Multiplying each side of (3) by $J^\perp$ and dividing by γ_{n+1}, we obtain:

$$\frac{\theta_n^\perp}{\gamma_{n+1}} = \alpha_{n+1} W_n^\perp \left(\frac{\theta_{n-1}^\perp}{\gamma_n} + Y_{n+1} \right) \tag{6}$$

where $\alpha_{n+1} := \gamma_{n+1}/\gamma_n$ and where, by row-stochasticity of W_n, we used $W_n^\perp \theta_n = W_n^\perp \theta_n^\perp$. Assumption 6b) ensures that Y_{n+1} given $\mathcal{F}_n$ is nearly distributed as $\mu_{\theta^\star \mathbf{1}}$ for large n, in the sense that $\mathbb{E}[f(Y_{n+1})|\mathcal{F}_n]$ converges to $\int f \mu_{\theta^\star \mathbf{1}}$ for any bounded continuous f. Recall also that by Assumption 3, $\lim_n \alpha_n = 1$. Loosely speaking, the update equation (6) is nearly a Markov chain with transition kernel $P^\star$ given by

$$P^\star(x, f) = \int \mathbb{E}\left[f(W_1^\perp (x + y)) \right] \, \mathrm{d}\mu_{\theta^\star \mathbf{1}}(y) \ ,$$

which represents the transition kernel of some homogeneous Markov chain $X_{n+1} = W_{n+1}^\perp (X_n + \xi_{n+1})$ where (ξ_n) is an i.i.d. process with distribution $\mu_{\theta^\star \mathbf{1}}$, $(W_n)_n$ is i.i.d. and W_{n+1} is independent from (X_n, ξ_{n+1}). Rigorous arguments will be given in an extended version of the paper; they rely on results for *generalized autoregressive model* [19].

Theorem 2. *Suppose Assumptions 1 to 6 and $\mathcal{L} = \{\theta^\star\}$. Then, $P^\star$ possesses an unique invariant measure $\pi^\star$ and almost surely, for any bounded continuous function f, $\lim_n n^{-1} \sum_{k=1}^n f(\theta_k^\perp / \gamma_{k+1})$ converges to $\pi^\star(f)$. Moreover,*

$$\lim_{n \to \infty} \gamma_n^{-1} \mathbb{E}(\theta_n^\perp) = R \cdot h^\star \tag{7}$$

$$\lim_{n \to \infty} \gamma_n^{-2} vec\, \mathbb{E}(\theta_n^\perp \theta_n^{\perp,T}) = S \cdot vec\,(T) \ , \tag{8}$$

where matrices R, S, T are given by:

$$R := (I_N - \mathbb{E}\left(W_1^\perp\right))^{-1} \mathbb{E}\left(W_1^\perp\right)$$
$$S := (I_{N^2} - \mathbb{E}\left(W_1^\perp \otimes W_1^\perp\right))^{-1} \mathbb{E}\left(W_1^\perp \otimes W_1^\perp\right)$$
$$T := Q^{\star,\perp} + R\, h^\star \cdot h^{\star,T} + h^\star \cdot h^{\star,T} R^T .$$

The first important point in Theorem 2 above lies in the convergence rate of the estimate. Sequence $\theta_n^\perp$ converges to zero at rate γ_n that is, faster than the standard convergence rate of standard SA algorithms. In practice, it means that agreement is achieved at overwhelming rate. Loosely speaking, the major part of the fluctuations of the estimation error $\theta_n - \theta^\star \mathbf{1}$ is contained in the consensus subspace (*i.e.* the linear span of $\mathbf{1}$) rather than the disagreement subspace. Theorem 2 also provides finer information about the asymptotic behavior of the normalized disagreement vector. The latter is asymptotically biased in the sense that it does not converge to zero in expectation. Our results allow to quantify the way individual estimates deviate from the average value, as a function of the gossip protocol and the local components h_i of the mean field h.

B. Average estimate

We need a last assumption on the mean field $\overline{h}$.

Assumption 7. *a) The mean field $\overline{h}$ is twice continuously differentiable in a neighborhood of $\theta^\star$.*

b) The derivative $H^\star := -\overline{h}'(\theta^\star)$ is strictly positive.

c) The step sizes satisfy $\log(\gamma_{n-1}/\gamma_n) = o(\gamma_n)$.

Note that the hypothesis $H^\star > 0$ is rather standard in the framework of SA when deriving second order results. When $\overline{h}$ is the gradient of an objective function f to be maximized, it is equivalent to the well-known second order sufficient condition which ensures that $\theta^\star$ is not only a critical point of f, but is also a local maximum. Assumption 7c) is for instance satisfied if $\gamma_n \sim \gamma_0 / n^a$ when $n \to \infty$, with $a \in (0.5, 1)$. Our result also extends to the case $a = 1$, but this generalization is postponed to an extended version of this paper in order to lighten the presentation.

Theorem 3. *Suppose Assumptions 1 to 7 and $\mathcal{L} = \{\theta^\star\}$. Then, the normalized average estimation error $\sqrt{\gamma_n}^{-1} \left(\overline{\theta}_n - \theta^\star\right)$ converges in distribution to*

$$\mathcal{N}\left(0, \sigma_{opt}^2 + \sigma_{com}^2\right)$$

where

$$\sigma_{opt}^2 = \frac{1}{2H^\star} \int \overline{y}^{\,2} \, \mu_{\theta^\star \mathbf{1}}(dy)$$

$$\sigma_{com}^2 = \frac{1}{2N^2 H^\star} \int \pi^\star(dx) \mu_{\theta^\star \mathbf{1}}(dy) \ (x + y)^T \cdots \times \mathrm{Cov}(W_1^{\perp,T} \mathbf{1}) \, (x + y) \, .$$

The proof is omitted due to the lack of space. It is mainly based on the results of [20], [21].

C. Discussion: The Impact of Non Double-Stochasticity

First, Theorem 3 states that the average estimation error converges to zero at rate $\sqrt{\gamma_n}$. This result was actually expected, as $\sqrt{\gamma_n}$ is the well-known convergence rate of standard SA algorithms. Second, Theorem 3 establishes that the variance of the asymptotic distribution of the normalized error is a sum

of two terms. The first term σ^2_{opt} is equal to the variance that would have been obtained in a fully centralized setting *i.e.*, in a scenario where all observations would be collected by a central processor running a SA algorithm. The second term σ^2_{com} characterizes the excess mean square error inherent to our distributed scenario. Now, note that $W_n^{\perp,T}\mathbf{1} = \mathbf{1}$ in the case where the gossip matrix W_n is column stochastic. In that case, $\text{Cov}(W_1^{\perp,T}\mathbf{1})$ is zero which implies that the excess variance σ^2_{com} is zero. Otherwise stated, the DSAA performs as well as a centralized algorithm whenever a doubly-stochastic gossip protocol is used. Reciprocally, $\sigma^2_{\text{com}} > 0$ whenever a non doubly-stochastic protocol is used, such as the simple broadcast scheme of Section II-C2.

V. Simulations

To show some simulation results, we consider the case of a complete graph with $N = 5$ nodes. The aim is to find the global minimum $\theta^\star$ of function $1/N\sum_{i=1}^N f_i(\theta)$, which is the average of agents' functions f_i. In our case, let us consider the following quadratic functions: $f_i(\theta) = a_i/2(\theta - c_i)^2$, where parameters a_i and c_i are associated to agent i. Parameter are chosen so that $\boldsymbol{c} = (-4, -3, -1, 2, 6)^T$ and $\boldsymbol{a} = \mathbf{1}$. In such a simple case the minimum corresponds to $\theta^\star = \frac{\bar{\boldsymbol{c}}}{\bar{\boldsymbol{a}}} = 0$. We simulate a a broadcast scheme with $\beta = 1/2$ and run $R = 1000$ independent realizations of the algorithm described in Section II-C using the decreasing step sequence $(1/n^{0.7})_{n\geq 1}$. A Gaussian noise sequence $(\xi_n)_{n\geq 1}$ with zero mean and variance σ^2 is added to give noisy observations $Y_{n,i} = -\theta_{n-1,i} + c_i + \xi_{n,i}$. The case $\sigma^2 = 0$ (absence of noise) is also considered. Figure 1 shows the histogram of the first component of $(\theta_n^\perp/\gamma_n)$ for n large enough $n = 30\,000$ in our case. As far as consensus is concerned, even without noise, pairwise scheme (with matrix W_n^P) has a worse behavior than broadcast. It can be quantitatively confirmed by our results (see Theorem 2). Indeed, in the broadcast case (with matrix W_n^B), the asymptotic variance of the sequence is zero thanks to the choice of $\beta = 1/2$ (in each trajectory $\lim_{n\to\infty}\theta_n^\perp/\gamma_n = \boldsymbol{c}$). However, from

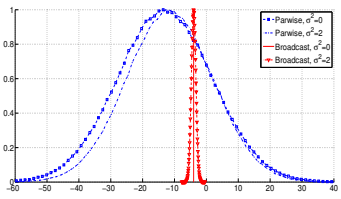

Figure 1. Histogram of the first component of the vector $\theta_n^\perp/\gamma_n$ for $n = 30\,000$ from 1000 independent runs for both protocols Pairwise/Broadcast under various noise condition.

the bias viewpoint, the situation is opposite. Figure 2 shows the histogram of the rescaled bias $\gamma_n^{-1/2}\left(\overline{\theta}_n - \theta^\star\right)$ for a large n ($n = 30\,000$ in our experiments). As the asymptotic variance of the sequence depends on the covariance matrix $\mathbb{E}\left[W_1^T\mathbf{1}\mathbf{1}^T W_1^T\right] - \mathbb{E}\left[W_1^T\mathbf{1}\right]\mathbb{E}\left[W_1^T\mathbf{1}\right]^T$ (see Theorem 3), it is zero for the pairwise scheme and non-zero for the broadcast (which is $N\beta^2 J^\perp$) .

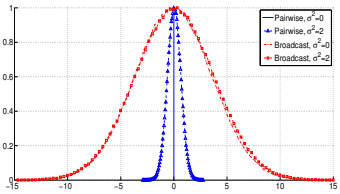

Figure 2. Histogram of the rescaled bias $\gamma_n^{-1/2}(\overline{\theta}_n - \theta^\star)$ for $n = 30\,000$ from 1000 independent runs for both protocols under various noise condition.

References

[1] A. Benveniste, M. Metivier, and Priouret P. *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] M. Rabbat and R. Nowak. Distributed Optimization in Sensor Networks. In *Proceedings of the 3rd international symposium on Information processing in sensor networks*, pages 20–27. ACM, 2004.

[4] A. Nedic and D.P. Bertsekas. Incremental Subgradient Methods for Non-differentiable Optimization. *SIAM Journal of Optimization*, 12(1):109–138, 2001.

[5] J. Lu, C.Y.I.K. Tang, P.R. Regier, and T.D. Bow. Gossip algorithms for convex consensus optimization over networks. *IEEE Trans. on Automatic Control*, 56(12):2917–2923, 2011.

[6] J. Tsitsiklis. *Problems in Decentralized Decision Making and Computation*. PhD thesis, Massachusetts Institute of Technology, 1984.

[7] J. Tsitsiklis, D. Bertsekas, and M. Athans. Distributed Asynchronous Deterministic and Stochastic Gradient Optimization Algorithms. *IEEE Trans. on Automatic Control*, 31(9):803–812, 1986.

[8] A. Nedic, A. Ozdaglar, and P.A. Parrilo. Constrained Consensus and Optimization in Multi-Agent Networks. *IEEE Trans. on Automatic Control*, 55(4):922–938, April 2010.

[9] S. Kar and J.M.F. Moura. Distributed consensus algorithms in sensor networks: Quantized data and random link failures. *IEEE Trans. on Signal Processing*, 58(3):1383–1400, 2010.

[10] S.S. Stankovic and, M.S. Stankovic, and D.M. Stipanovic. Decentralized Parameter Estimation by Consensus Based Stochastic Approximation. *IEEE Trans. on Automatic Control*, 56(3):531–543, march 2011.

[11] A. Nedic and A. Ozdaglar. Distributed Subgradient Methods for Multi-Agent Optimization. *IEEE Trans. on Automatic Control*, 54(1):48–61, 2009.

[12] P. Bianchi, G. Fort, and W. Hachem. Performance of a Distributed Stochastic Approximation Algorithm. *IEEE Trans. on Inform. Theory (submitted) arXiv1203.1505 [math.OC]*, 2012.

[13] S.S. Ram, A. Nedic, and V.V. Veeravalli. Distributed Stochastic Subgradient Projection Algorithms for Convex Optimization. *Journal of optimization theory and applications*, 147(3):516–545, 2010.

[14] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah. Randomized Gossip Algorithms. *IEEE Trans. on Inform. Theory*, 52(6):2508–2530, 2006.

[15] T.C. Aysal, M.E. Yildiz, A.D. Sarwate, and A. Scaglione. Broadcast Gossip Algorithms for Consensus. *IEEE Trans. on Signal Processing*, 57(7):2748–2761, 2009.

[16] A. Nedic. Asynchronous Broadcast-Based Convex Optimization Over a Network. *IEEE Trans. on Automatic Control*, 56(6):1337–1351, june 2011.

[17] P. Bianchi and J. Jakubowicz. On the convergence of a projected multi-agent stochastic gradient algorithm for non-convex optimization. *IEEE Trans. on Autom. Control, to appear*, 2013.

[18] F. Bénézit. *Distributed Average Consensus for Wireless Sensor Networks*. PhD thesis, EPFL, 2009.

[19] M. Duflo. *Algorithmes stochastiques*. Springer Berlin, 1996.

[20] G. Fort, E. Moulines, and P. Priouret. Convergence of Adaptive and Interacting Markov chain Monte Carlo algorithms. *Ann. Statist.*, 39(6):3262–3289, 2012.

[21] G. Fort. Central limit theorems for stochastic approximation algorithms. Technical report, 2012.