Distributed Stochastic Approximation: Weak Convergence and Network Design

Miloš S. Stanković, Nemanja Ilić and Srdjan S. Stanković

December 11, 2014

Abstract

This paper presents a study of two characteristic distributed stochastic approximation algorithms based on broadcast gossip on communication networks represented by digraphs. Weak convergence of these algorithms is proved based on the formulation of a limit ordinary differential equation (ODE), which is a function of communication rates and network parameters. Using these results, a general methodology is proposed for network design aimed at achieving the desired asymptotic properties of the estimates at consensus. Convergence rate of the algorithm is analyzed and further improved using an attached stochastic differential equation. Some simulation results give an illustration of the theoretical concepts.

1 Introduction

Stochastic approximation (SA) algorithms have been in the focus of researchers for more than sixty years, e.g., [1–3] and the references therein. More than thirty years ago a distributed SA algorithm based on a first-order consensus scheme was proposed in [4] and analyzed in detail for both constant and tapering step-sizes [4–8]. Recently, there has been a growing interest for distributed algorithms, networked control systems and multi-agent systems. As a consequence, new contributions to the distributed SA

*Innovation Center, School of Electrical Engineering, University of Belgrade, Serbia (e-mail: milos.stankovic@ic.etf.rs).
†School of Electrical Engineering, University of Belgrade, Serbia (e-mail: nemili@etf.rs).
‡School of Electrical Engineering, University of Belgrade, Serbia (e-mail: stankovic@etf.rs).

This work was supported by the EU Marie Curie CIG (PCIG12-GA-2012-334098) and the Serbian Ministry of Education, Science and Technological Development.
algorithms have appeared, motivated by numerous applications in optimization, estimation, detection and control in which decentralization or distribution of functions is either unavoidable, dictated by information structure constraints, or beneficial, aimed at parallelizing computations, increasing robustness, etc. [9–16]. The problem setting in all these approaches follows essentially the main line of thought from [4, 6, 7]. However, depending on the context, specific, often more restrictive assumptions have been adopted, especially in relation with the communication scheme aimed at achieving consensus between the agents performing local computations. In [17] fixed communication links are considered in relation with the LMS estimation algorithm. In [9–11, 18] random communication links are assumed in relation with distributed optimization algorithms, which can be formally considered as specific SA algorithms. However, in all these approaches the assumption about double stochasticity of the random communication (gossip) matrices has been adopted. In [12, 14, 15] this assumption has been relaxed, requiring double stochasticity of the mean value of these matrices.

In this paper we shall approach distributed SA algorithms from a more general standpoint, assuming that the underlying multi-agent network is represented by a strongly connected directed graph with arbitrary topology. In particular, we shall consider two characteristic SA algorithms based on asynchronous broadcast gossip [12, 19]. Starting from the basic papers [4, 7], we shall first present a new insight into the applied gossip scheme and prove its convergence with probability one (w.p.1) (generalizing the result of [19]). Based on this result, we shall propose two distinct schemes for communication network design (calculation of transmitting probabilities and communication gains), based on the desired properties at consensus. Then we shall present, extending the results of Kushner and Yin [7], weak convergence results for the analyzed SA algorithms, focusing on the derivation of the limit ordinary differential equation (ODE), which is a function of the network parameters. Besides its theoretical value, this result is important for practice, since it enables application of the presented network design methodology to obtaining the desired limit behavior. It is to be stressed that the same limit ODE is obtained for tapering step sizes (typical for SA algorithms). In addition, it is shown that the normalized asymptotic error can be modeled by a stochastic differential equation (SDE) and demonstrated how it can be utilized for improving the convergence rate of the algorithms. The given simulation results serve as an illustration of the derived theoretical results.

2 Problem formulation and algorithms

Consider a network of $N$ agents, represented by a directed graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N} = \{1, \ldots, N\}$ is the set of nodes, and $\mathcal{E} = \{(i, j)\}$ the set of directed links from node $i$ to node $j$. Let $A_\mathcal{G}$ be the adjacency matrix
of $\mathcal{G}$. Define the sets of out-neighbors and in-neighbors of the $i$-th node $\mathcal{N}^o_i = \{ j \in \mathcal{N} | (i, j) \in \mathcal{E} \}$ and $\mathcal{N}^i_i = \{ j \in \mathcal{N} | (j, i) \in \mathcal{E} \}$, respectively. Assume that each agent has an internal clock that ticks independently from the other clocks according to a rate $\mu_i > 0$ Poisson process, $i = 1, \ldots, N$. At each tick of the internal clock, the node $i$ broadcasts its current state to the nodes $j \in \mathcal{N}^o_i$; due to possible link failures, some neighbors may not receive this message. When successfully received, the message leads to an update of the states of the neighboring nodes, while the remaining nodes from $\mathcal{G}$ preserve their previous states. Such a type of communication scheme belongs to the class of broadcast gossip algorithms, see e.g. [19, 20] and the references therein. We shall replace in our analysis the set of local clocks by a global virtual clock with the rate $\mu = \sum_i \mu_i$ that ticks whenever any local clock ticks; the $n$-th tick of the global clock is considered as the discrete time instant (iteration) $n$. At any instant $n$, let $p_i > 0$ be the probability for $i$-th clock to tick, $J^i(n) \subseteq \mathcal{N}^o_i$ the set of nodes that received the message from node $i$, $p_{ij} = P\{ j \in J^i(n) \}$ the probability of $j$-th node to receive the message from node $i$ and $\bar{d}_j = \sum_{i \in \mathcal{N}_j^i} p_{ij} p_i$ the probability of $j$-th node to receive a message and update its state.

Let the $i$-th clock ticks at the instant $n$ and let $X^i_n \in \mathbb{R}^p$ be the current state of the $j$-th agent, $j = 1, \ldots, N$. The agents are supposed to generate their new states in the following two characteristic ways:

**Algorithm update-convexify (AUC).** At the first step, the agent $i$ calculates its updated state

$$\hat{X}^i_{n+1} = X^i_n + \varepsilon f^i(X^i_n, \xi^i_n),$$

(1)

and sends it to its neighbors, where $f^i(X^i_n, \xi^i_n)$ is the "observation" of agent $i$ at time $n$, $\xi^i_n$ is the noise and $\varepsilon > 0$ the step-size. The neighbors $j \in J^i(n)$ generate at the first step their own updated states $\hat{X}^j_{n+1}$ using (1), and calculate, at the second step, their new states as convex combinations

$$X^j_{n+1} = \beta_{ij} \hat{X}^j_{n+1} + (1 - \beta_{ij}) \hat{X}^i_{n+1},$$

(2)

where $0 < \beta_{ij} < 1$, while $X^i_{n+1} = \hat{X}^i_{n+1}$ and $X^k_{n+1} = X^k_n$ for $k \notin J^i(n)$, $k \neq i$.

**Algorithm convexify-update (ACU).** The agent $i$ sends at instant $n$ its current state $X^i_n$ to the neighboring nodes. At the first step, the agents $j \in J^i(n)$ calculate convex combinations

$$\hat{X}^j_n = \beta_{ij} X^j_n + (1 - \beta_{ij}) X^i_n,$$

(3)

and, at the second step, update their states using

$$X^j_{n+1} = \hat{X}^j_n + \varepsilon f^j(\hat{X}^j_n, \xi^j_n);$$

$$X^k_{n+1} = X^k_n$$

for all $k \notin J^i(n)$. 3
AUC has been described in [14][16][21], and ACU in [12]; in [17] both schemes are presented in a deterministic context.

Defining $X_n = (X_n^1, \ldots, X_n^N)$ and $\xi_n = (\xi_n^1, \ldots, \xi_n^N)$ (the notation $(x_1, \ldots, x_N)$ is used throughout the paper to represent the column vector obtained by concatenating vectors $x_1, \ldots, x_N$), we obtain from (1) and (2) the following compact representation for both AUC and ACU:

$$X_{n+1} = \tilde{A}_nX_n + \varepsilon\tilde{C}_nF(Y_n, \xi_n)$$

(5)

where:
- $\tilde{A}_n = A_n \otimes I_p$, where $A_n = [a_{jk}(n)]$ is a random matrix defined in such a way that
  $$a_{jk}(n) = \begin{cases} 1, & j \notin J_i(n), k = j \\ \beta_{ij}, & j \in J_i(n), k = j \\ 1 - \beta_{ij}, & j \in J_i(n), k = i \\ 0 & \text{elsewhere} \end{cases}$$

(0 < $\beta_{ij}$ < 1) ($i,j,k = 1,\ldots, N$);
- for AUC: $\tilde{C}_n = A_n D_n \otimes I_p$, in which $D_n = \text{diag}\{d_1(n), \ldots, d_N(n)\}$, where $d_k(n)$ is a binary random variable equal to one if $k \in J_i(n) \cup \{i\}$ and zero otherwise;
- for ACU: $\tilde{C}_n = D_n^- \otimes I_p$, where $D_n^-$ has the same elements as $D_n$, except $d_i(n)$ which is equal to zero;
- $F(X, \xi) = (f_1(X^1, \xi^1), \ldots, f_N(X^N, \xi^N))$, where $X = (X^1, \ldots, X^N)$ and $\xi = (\xi^1, \ldots, \xi^N)$, so that $F(X_n, \xi_n) = F(X, \xi)|_{x=X_n, \xi=\xi_n}$;
- for AUC: $Y_n = X_n$; for ACU: $Y_n = \tilde{A}_nX_n$.

Let $\mathcal{F}_n$ be an increasing sequence of $\sigma$-algebras such that $\mathcal{F}_n$ measures $\{X_i, i \leq n; \xi_i, A_i, i < n\}$ and let $E_k\{\cdot\}$ denote $E\{\cdot|\mathcal{F}_k\}$.

3 Broadcast Gossip Communication Scheme

3.1 Convergence to Consensus

In this subsection we shall pay attention to the properties of the communication part of the above given algorithm, represented by

$$z_{n+1} = A_nz_n$$

(7)

where $z_n \in R^N$, $|z_0| < \infty$ and $A_n$ is defined above ($\|A\|$ denotes the infinity norm of a matrix $A$). Define $\Phi(n/k) = A_n \cdots A_k$ for $n \geq k$, $\Phi(n|n + 1) = I_N$, so that $z_{n+1} = \Phi(n,0)z_0$. Let $\tilde{z}_n = E\{z_n\}$; we have $\tilde{z}_n = E\{\Phi(n - 1|0)\}z_0 = \tilde{A}^n\tilde{z}_0$, where $\tilde{A} = E\{A_n\} = \sum_i A_i^0\pi_i$, in which $A_i^0$ denotes the $l$-th realization of $A_n$ and $\pi_i$ its probability (which follows from $p_i$ and $p_{ij}$, $i,j = 1, \ldots, N$).

(A.1) The digraph $G$ is strongly connected [22].

The following lemma slightly generalizes Lemma 2.1 from [7] and applies it to AUC and ACU.
Lemma 1 Let (A.1) hold. Then:

a) \( \lim_{n \to \infty} \bar{z}_n = c1 \), where \( \bar{c} = \bar{\phi}z_0 \) and \( \bar{\phi} = (\bar{\phi}_1, \ldots, \bar{\phi}_N)^T \) satisfies \( \bar{\phi} \bar{A} = \bar{\phi} \) (1 = (1, \ldots, 1));

b) \( \Phi_k = \lim_{n \to \infty} \Phi(n|k) \) exists w.p.1 and its rows are all equal; moreover,

1) \( \mathbb{E}\{|\Phi(n|k) - \Phi_k|\} \to 0 \) geometrically as \( n - k \to \infty \), uniformly in \( k \);

2) \( \mathbb{P}\{\lim_{n \to \infty} z_n = c1\} = 1 \) for some random variable \( c \).

Proof: Assertion a) follows from (A.1), which implies that the adjacency matrix \( A_G \) is primitive [22,23]; consequently, \( \bar{A} \) is also primitive and, therefore, \( \bar{\Phi} = \lim_{n \to \infty} \mathbb{E}\{\Phi(n|k)\} = \lim_{n \to \infty} \bar{A}^{n-k} \) exists for any fixed \( k \) and has the form \( \bar{\Phi} = \left[ \bar{\phi}^T \cdots \bar{\phi}^T \right]^T \) where \( \bar{\phi} = (\bar{\phi}_1, \ldots, \bar{\phi}_N)^T \), \( \bar{\phi}_i > 0 \), \( i = 1, \ldots, N \).

In the context of the described broadcast gossip schemes, there exist a scalar \( p_0 > 0 \) and an integer \( m_0 > 0 \) such that for all \( n \) the probability that the node \( i \) communicates to any node \( j \in N_i \) on \( [n, n+m_0) \) is greater than or equal to \( p_0 \). Therefore, there exist \( \alpha'_0 > 0 \) and an integer \( m' > 0 \) such that the elements of \( \Phi(n+m'|n) \) having the same indices as the nonzero elements of the adjacency matrix \( A_G \) are all \( \geq \alpha'_0 \) w.p.1 for all \( n \). Also, by (A.1), \( \Phi(n+m'|n) \) is primitive, implying that there exist \( \alpha''_0 > 0 \) and an integer \( m'' > 0 \) such that the components of \( \Phi(n+m''|n) \) are all \( \geq \alpha''_0 \) w.p.1 for all \( n \) [16]. As a consequence, there are \( \alpha_0 > 0 \) and an increasing sequence of finite w.p.1 random times \( \nu_i \) such that the components of \( \Phi(\nu_{i+1}|\nu_i) \) are all \( \geq \alpha_0 \) w.p.1. Using this fact and the result of Lemma 2.1 from [7], we come to the conclusion that \( \Phi(n|k) \) converges w.p.1 when \( n \to \infty \) to a row stochastic matrix \( \Phi_k \) whose rows are all equal, i.e., \( \Phi_k = \left[ \phi(k)^T \cdots \phi(k)^T \right]^T \), with \( \phi(k) = (\phi_1(k), \ldots, \phi_N(k))^T \). Following directly the proof of Lemma 2.1 from [7] we can prove assertion b1); coming back to [7], we conclude that \( \lim_{n \to \infty} z_n = \lim_{n \to \infty} \Phi(n - 1|0)z_0 = \Phi_0z_0 = c1 \) w.p.1, where \( c = \phi(0)z_0 \). Thus the result.

Remark 1 The above results represent a generalization of the results from [19], where the same broadcast gossip scheme is analyzed, assuming \( p_i = p_j = 1/N, \beta_{ij} = \beta, i, j = 1, \ldots, N \) and \( 1^T A = 1^T \); it has been proved in [19] that in this case \( \mathbb{E}\{c\} = \frac{1}{N}1^Tz_0 \), i.e., \( \bar{\phi} = \frac{1}{N}1^T \). The assertion a) defines an important connection between the network properties and vector \( \bar{\phi} \) which will be utilized below.

3.2 Network Design

The above analysis opens up a possibility to design the network in accordance with its desired asymptotic behavior. Having in mind the nature of communications protocol and the above results, it is possible to pose the following problem:
Let \( \phi = (\phi_1, \cdots, \phi_N)^T \), \((0 \leq \phi_i < 1, \sum_j \phi_j = 1)\), be given; for a given structure of digraph \( G \) satisfying (A.1), find parameters \( p_i > 0, i = 1, \ldots, N, \sum_i p_i = 1, \) and \( \beta_{ij}, (0 < \beta_{ij} < 1), i,j = 1, \ldots, N, \) such that \( \lim_{n \to \infty} A^{n-k} = \begin{bmatrix} \phi^T & \cdots & \phi^T \end{bmatrix}^T \) for any fixed \( k \).

Formally, we have to solve, under the given constraints, the equation \( \bar{\phi} \bar{A} = \bar{\phi} \) for the unknown parameters \( p_i \) and \( \beta_{ij} \) figuring in \( \bar{A} \), \( i,j = 1, \ldots, N \). Assume first, for the sake of exposition clarity, that \( J_i(n) = N_i^o \). If \( A^{[i]} \) is the realization of \( A_n \) (defined by (6)) corresponding to a tick of the \( i \)-th clock, we have then the basic relation \( \bar{\phi} \sum_i A^{[i]} p_i = \bar{\phi} \). Having in mind the large number of unknowns, we shall consider two special cases: a) parameters \( \beta_{ij} \) are a priori fixed and we are looking for \( N \) transmission probabilities \( p_i \); b) probabilities \( p_i \) are fixed, and we are looking for the parameters \( \beta_{ij} \).

Case a). From \( \bar{\phi} \sum_i A^{[i]} p_i = \bar{\phi} \) we obtain \( \bar{A}^T p = 0 \), where

\[
\bar{A}^L = \begin{bmatrix}
\sum_j \bar{\phi}_j \beta_{1j} & \cdots & -\bar{\phi}_1 \beta_{N1} \\
-\bar{\phi}_2 \beta_{12} & \sum_j \bar{\phi}_j \beta_{2j} & \cdots & \vdots \\
\vdots & \ddots & \ddots & \vdots \\
-\bar{\phi}_N \beta_{1N} & \cdots & -\sum_j \bar{\phi}_j \beta_{Nj}
\end{bmatrix}
\]

and \( p = (p_1, \cdots, p_N) \) (implicitly, \( \beta_{ij} = 0 \) for \( j \notin N_i^o \)). As it is possible to observe, \( (A^L)^T \) is in the form of a weighted Laplacian of the given graph \( G \); therefore, \( \text{rank}(\bar{A}^L) = n - 1 \) under (A.1). Any \( N \times N \) matrix \( \bar{A}^p, \) formed from any set of \( N - 1 \) linearly independent rows from \( \bar{A}^L \) and the row \( N \)-vector \( 1^T \) (coming out form the additional basic relation \( \sum_i p_i = 1 \)), is nonsingular (it is easy to verify that \( 1^T \) is linearly independent of the rows of \( \bar{A}^L \)). Therefore, we have the set of linear equations \( \bar{A}^p p = (0 \cdots 0 1)^T \), which uniquely determines vector \( p \). It is easy to check that the solution \( p \) satisfies \( 0 < p_i < 1 \) for all admissible \( \beta_{ij}, i,j = 1, \ldots, N \).

Case b). Assume first that \( \beta_{ij} = \beta_i \) for all \( j, i = 1, \ldots, N \). Then, the equation \( \bar{\phi} \bar{A} = \bar{\phi} \) can be written as \( \bar{A}^{EL} \beta = 0 \), where

\[
\bar{A}^{EL} = \begin{bmatrix}
p_1 \sum_{j,j\neq 1} \bar{\phi}_j & -p_2 \bar{\phi}_1 & \cdots & -p_N \bar{\phi}_1 \\
-p_1 \bar{\phi}_2 & p_2 \sum_{j,j\neq 2} \phi_j & \cdots & -p_N \bar{\phi}_2 \\
-p_1 \bar{\phi}_N & \cdots & p_N \sum_{j,j\neq N} \bar{\phi}_j
\end{bmatrix}
\]
and $\beta = (\beta_1, \ldots, \beta_N)$. Under (A.1), $(\bar{A}^{EL})^T$ has the form of a weighted Laplacian of the graph $\mathcal{G}$, satisfying $\text{rank}\{\bar{A}^{EL}\} = n - 1$. It is easy to conclude that there is an infinite number of solutions for $\beta$ satisfying $0 < \beta_i < 1$, $i = 1, \ldots, N$. In general, when $\beta_{ij}$ are different for different $j$, it is straightforward to conclude that the desired solution exists, but with more degrees of freedom.

In the general case of communication outages (when $\mathcal{J}(n) \neq N_i^c$) a similar analysis can be done in a straightforward way, taking into account transmission probabilities $p_{ij}$ (assumed to be given).

**Remark 2** According to the above results, consensus averaging, providing $\bar{\phi} = 1/N 1^T$, can be achieved for any given network structure by choosing either of the two approaches (it is not necessary to assume bi-directional communications, as in [12], for example).

### 4 Distributed Stochastic Approximation

#### 4.1 Weak Convergence

Weak convergence of the algorithm [15] will be analyzed under the following additional assumptions:

(A.2) $\{\xi_n\} = (\eta_n, \zeta_n)$, where $\{\eta_i\}$ is a sequence of bounded random variables and $\{\zeta_i\}$ a random sequence with zero mean and finite fourth moment;

(A.3) Sequences $\{A_n\}$ and $\{\xi_n\}$ are independent;

(A.4) Let $F(X, \xi) = F_1(X, \eta) + F_2(X)\zeta$, where both $F_1(\cdot, \cdot)$ and $F_2(\cdot)$ are continuous; then, $\lim_{n \to \infty} E_k\{F_1(X, \eta_n)\} = \bar{F}(X)$, where $\bar{F}(X) = (\bar{f}^1(X), \ldots, \bar{f}^N(X))$ is a continuous function and $E_k\{\zeta_n\} \to 0$ in probability when $n - k \to \infty$.

(A.5) Let $\bar{D} = D/\{D_n\} = \text{diag}\{\bar{d}_1, \ldots, \bar{d}_N\}$, $\bar{F}(x)$ denote $\bar{F}(x, \ldots, x)$ and $F(x, \xi)$ denote $F(x, \ldots, x, \xi)$, where $x = (x_1, \ldots, x_p)$ is a dummy variable. Then the ODE

$$
\begin{align*}
\dot{x}_1 &= \bar{\phi}_1 \bar{d}_1 \bar{f}_1^1(x) + \cdots + \bar{\phi}_N \bar{d}_N \bar{f}_1^N(x) \\
\vdots \quad & \quad \quad \vdots \\
\dot{x}_p &= \bar{\phi}_1 \bar{d}_1 \bar{f}_p^1(x) + \cdots + \bar{\phi}_N \bar{d}_N \bar{f}_p^N(x)
\end{align*}
$$

has a unique solution for each initial condition, where $\bar{f}_i^j(x)$ is the $i$-the component of $\bar{f}^j(x)$, $i = 1, \ldots, p$, $j = 1, \ldots, N$.

Let $\Phi(n|k) = \Phi(n|k) \otimes I_p$, $\bar{\Phi}_k = \Phi_k \otimes I_p$, $\Psi_k = \Psi_k \otimes I_p$, with $\bar{\Phi}(n|k) = \bar{\Phi}(n|k)D_k$ and $\Psi_k = \Phi_k D_k$ for AUC, and $\bar{\Psi}(n|k) = \Psi(n|k + 1)D_k$ and $\Psi_k = \Phi_k D_k$ for ACU. Following strictly [7], let $\{n_\varepsilon\}$ be a sequence tending to $\infty$ when $\varepsilon \to 0$, such that $\sqrt{\varepsilon} n_\varepsilon \to 0$ and $\sup_k P\{|\Phi(k + n_\varepsilon|k) - \Phi_k| \geq \varepsilon^2\} \leq \varepsilon^2$. 

7
Let $X^\varepsilon(t) = X_n$ for $t \in [(n - n_\varepsilon) \varepsilon, (n - n_\varepsilon + 1) \varepsilon]$ for $n \geq n_\varepsilon$, where $X_n$ is defined in (5); also, let $X^\varepsilon_0 = \Phi(n\varepsilon(0))X_0 + \varepsilon \sum_{k=0}^{n_\varepsilon - 1} \Psi_k F(Y_k, \xi_k)$.

**Theorem 1** Assume (A.1) – (A.5). Let $X(0) = \lim_{\varepsilon \to 0} X^\varepsilon_0 = (x_0, \ldots, x_0)$. Then for both AUC and ACU $X^\varepsilon(t)$ is tight in $D[0, \infty)$ and converges weakly to $X(t) = (x(t), \ldots, x(t))$, $t \in \mathbb{R}$, where $x(t)$ satisfies (5) with initial condition $x_0$ ($D[0, \infty)$ denotes the space of vector valued functions on $[0, \infty)$ which are right continuous and have left hand limits, with the Skorokhod topology [3, 7]).

**Proof:** The proof is based on Theorem 3.1 in [7]. Iterating (5) back to the initial condition, we obtain

$$X_{n+1} = X^\varepsilon_0 + \varepsilon \sum_{k=n_\varepsilon}^{n} \Psi_k F(Y_k, \xi_k) + \varepsilon \tilde{\Phi}_n + \varepsilon \Phi(n\varepsilon) - \Phi(n\varepsilon(0))X_0, \quad (9)$$

where $\tilde{\Phi}_n = \sum_{k=n_\varepsilon}^{n} \tilde{\Psi}(n\varepsilon)_k - \Psi_k F(Y_k, \xi_k)$. Following [3, 7] we assume that $\{X_n\}$ is bounded and obtain that $\sup_{\varepsilon, n} E\{|\tilde{\Phi}_n\varepsilon^3\} < \infty$ and $\sup_{\varepsilon, n \geq n_\varepsilon} E\{|X_n+1 - X_n|/\varepsilon^2\} < \infty$, implying that $\{X^\varepsilon(t)\}$ is tight in $D[0, \infty)$ and that all limit paths are Lipschitz continuous for both AUC and ACU.

At the second step, following further [7], we introduce

$$M_g(t) = g(X(t)) - g(X(0)) + \int_0^t g_X(X(s))(\dot{\Phi} \tilde{D} \otimes I_p) \tilde{F}(X(s))ds, \quad (10)$$

and show that $M_g(t)$ is a martingale for any real valued function $g(\cdot)$ with compact support and continuous second derivatives. As $M_g(t)$ is Lipschitz continuous for both AUC and ACU, this implies that $M_g(t)$ is constant and equal to zero; consequently, $X^\varepsilon = \Phi \tilde{D} \otimes I_p F(X)$. Furthermore, as the rows of $\Phi \tilde{D}$ are equal, all the $p$-vector components of $X(t)$ must be equal, implying further that $X^\varepsilon(t)$ converges weakly to $X(t) = (x(t), \ldots, x(t))$, where $x(t)$ satisfies the ODE (3).

Notice that the results in [7] are related to a version of (9) in which $\Psi(n\varepsilon) = \Phi(n\varepsilon + 1)$ and $\Psi_k = \Phi_k + 1$ and $Y_k = X_k$, and in which all nodes update their states at every node $n$. For an extension of these results to AUC and ACU it is essential, at this point, to demonstrate that

$$B^\varepsilon(t) \xrightarrow{\varepsilon \to 0} g_X(X(t))(\Phi \tilde{D} \otimes I_p) \tilde{F}(X(t)), \quad (11)$$

where $B^\varepsilon(t) = g^\varepsilon_X(X^\varepsilon(t))E_{l+n_\varepsilon} \{\tilde{\Phi}_l \tilde{D}_l \Phi E\{A_l D_l \pi_l \}, l \leq t < (l+1) \varepsilon$ (compare with [7], proof of Theorem 3.1, Eqs. (3.6) – (3.10)). Indeed, for AUC we have $E\{\Psi_k \} = E\{\Phi_k \tilde{D} \} = \Phi E\{A_k \tilde{D} \} = \Phi \sum_{l} A[l] \tilde{D}[l] $, where $\tilde{D}[l]$ are realizations of $D_n$ connected to the realizations $A[l]$; moreover, $E\{\Psi_k \} = \Phi(\tilde{A} - \sum_{l} A[l] D[l] \pi l)$, where $D[l] = I - D[l]$. As $\Phi \tilde{A} = \Phi$ and $A[l] D[l] = D[l]$, having in mind the structure of $A[l]$, one concludes that
and reducing the stationary covariance of $u$ of convergence of the algorithms, but also for network design aimed at

$$E\{\Psi_k\} = \Phi(\tilde{A} - D^-) = \Phi \tilde{D},$$
where $D^- = \sum_l D[l-\pi_l]$ is composed of node probabilities of not updating the state at a given instant $n$. For ACU, $E\{\Psi_k\} = \Phi \tilde{D}$, having in mind that then $A_k+1$ and $D_k$ are independent (in [7], $E\{\Psi_k\} = \tilde{\Phi}$). Note that the numerical values of $\tilde{D}$ are not the same for AUC and ACU. Note also that convexification of $X_n$ in ACU does not influence the main conclusion, having in mind that again $X^\varepsilon(t) \to X(t)$ uniformly on bounded intervals and $X(t) = (x(t), \ldots, x(t))$ as a consequence of the structure of $\Phi \tilde{D}$.

4.2 Asymptotics for large $t$ and small $\varepsilon$

(A.5) Let (8) have a unique, in the sense of Lyapunov, stable point $x^*$ which is globally attracting.

Theorem 2 Let the conditions of Theorem 1 and (A.5) be satisfied. Assume that the set $\{X^\varepsilon(t), t \geq 0, \varepsilon > 0\}$ is tight. Then $X^\varepsilon(t)$ converges weakly to $X^* = (x^*, \ldots, x^*)$ as $t \to \infty$ and $\varepsilon \to 0$.

Proof: The proof follows from Theorem 1; it is analogous to Theorem 5.1 in [7].

Sharper bounds on the asymptotic normalized error $U^\varepsilon_n = \frac{X_n - X^*}{\sqrt{\varepsilon}}$ can be obtained, assuming that $U^\varepsilon_n$ is tight and that, additionally, $E\{F(X^*, \xi_n)\} = 0$ [7]. Define $W^\varepsilon_n = \sqrt{\varepsilon} \sum_{k=M+n+1}^{N} \Psi_k F(Y_k, \xi_k)$ where $M$ is large enough. Applying again the results from [7], it is possible to demonstrate for both AUC and ACU that, under additional assumptions (including independence of $\xi_i^\varepsilon$, $i = 1, \ldots, N$), $U^\varepsilon_n$ and $W^\varepsilon_n$ converge weakly to $U(t) = (u(t), \ldots, u(t))$ and $W(\cdot) = (w(t), \ldots, w(t))$, respectively, where $u(t)$ and $w(t)$ satisfy the following linear stochastic differential equation

$$du = J ud t + dw,$$

in which $J = ((\Phi \tilde{D} \otimes I_p) \tilde{F}(x^*))_x$ is the Jacobian of $(\Phi \tilde{D} \otimes I_p) \tilde{F}(x)$ at $x = x^*$ and $w(\cdot)$ is a $p$-dimensional Wiener process with covariance

$$\text{cov}w(1) = \sum_{i=1}^{N} E\{\phi_i(k)^2 d_i(k)^2\} R_i,$$

where $R_i = \text{cov} f^i(x^*, \xi_i^\varepsilon)$.

Relations (12) and (13) can be used not only for estimating the rate of convergence of the algorithms, but also for network design aimed at reducing the stationary covariance of $u$ in (12) by minimizing the upper bound of $\|\text{cov} w(1)\|$. Having in mind that $E\{\phi_i(k)^2 d_i(k)^2\}$ is not easy to calculate, one can formulate the following practical criterion to be minimized:

$$\sum_{i=1}^{N} \tilde{d}_i^2 \|R_i\|^{-1} \sum_k \tilde{d}_k^2 \|R_k\|^{-1}.$$  

The optimal value of $\tilde{\phi}_i$ satisfying $\sum_i \tilde{\phi}_i = 1$ is $\tilde{\phi}_i^* = \tilde{d}_i^{-2} \|R_i\|^{-1} / \sum_k \tilde{d}_k^{-2} \|R_k\|^{-1}$. Concrete network parameters can now
be found by using one of the two proposed methodologies in Section III, starting from the obtained $\bar{\phi}_i^* \ i = 1, \ldots, N$; notice that the assumption $E\{F(X^*, \xi_n)\} = 0$ implies that the convergence point at consensus does not depend on the network parameters, see (8).

Remark 3 The above given analysis can be extended to the case when $\varepsilon$ is replaced by a time varying sequence $\{\varepsilon_n\}$, such that $\varepsilon_n > 0$, $\varepsilon_n \to 0$ and $\sum \varepsilon_n = \infty$, as it is common in SA procedures [1,3]. Starting from the assumptions of Theorem 2, one can readily apply the methodology from [3] and obtain that the same ODE as above characterizes the limit paths of the estimates; moreover, it is possible to show weak convergence to $x(\cdot) = x^*$, under appropriate stability assumptions. The w.p.1 convergence results can also be obtained assuming that $\sum \varepsilon_n^2 < \infty$. Details related to the tapering step-size are not in the focus of this paper; however, it is to be emphasized that the ODE (8) derived the above arguments is still applicable, together with the whole proposed network design methodology (see some simulation in Section V).

Remark 4 In [12] an asynchronous algorithm of ACU type for broadcast-based convex minimization of sum of local criteria has been analyzed assuming bidirectional links. In [14] a SA algorithm of AUC type is treated under a restrictive assumption that $\|E\{A_n^T(I - 11^T)A_n\}\| < 1$; analogous conditions are not satisfied, in general, under the above adopted assumptions for the proposed algorithms, but still the convergence to consensus has been proved.

4.3 Network Design

The primary aim of the network design for (5) is to provide the desired convergence point at consensus defined by the condition $\dot{x} = 0$ in the main ODE (8). From this point of view, the basic relationship between the network parameters $p_i$ and $\beta_{ij}$, on one side, and $\bar{\phi}_k$, on the other, has been discussed in subsection III.B. In addition to this, ODE (8) incorporates the updating probabilities of the nodes $\bar{d}_i$, $i = 1, \ldots, N$ (which can be easily calculated using $p_i$ and $p_{ij}$). Assume that the designer’s goal is to obtain a predefined set of weights $w_i = \bar{\phi}_i \bar{d}_i$, $w_i > 0$, $i = 1, \ldots, N$ in (8); then, the following network design procedure is proposed:

1) adopt a set of transmission probabilities $p_i > 0$ and define $\bar{d}_i$, $i = 1, \ldots, N$, and

2) design the network by using the second methodology from III.B (Case b), by choosing such communication gains $\beta_{ij}$ that the set $\bar{\phi}_k = w_k / \bar{d}_k$, $k = 1, \ldots, N$ is obtained at consensus (after appropriate normalization).

If, for example, the proposed SA algorithms are aimed at distributed minimization of a given criterion $J = \sum_{i=1}^N w_i E\{J_i(x, \xi_n)\}$, $w_i > 0$, then $w_i$ are predefined weights; in this case $f^i(X^i, \xi_n^i)$ in (1) and (4) are gradients or
pseudo-gradients of $J_i(x, \xi_n^i)$ at $x = X^i$ (see e.g. \cite{10,12,18}). Obviously, the case of equal weights (consensus averaging) can be obtained for any given network topology satisfying (A.1) (much more restrictive assumptions are adopted in \cite{9,11,12,18}).

Moreover, it is to be noticed that there exists an additional possibility for obtaining the desired asymptotic behavior of (8) by choosing different gains $\varepsilon_i = v_i \bar{\varepsilon}$ for different agents. We have then in (8) coefficients of the form $w_i = v_i \bar{\varepsilon}_i$, in which $v_i$ allows a direct “compensation” of any given $\bar{d}_i$, so that $w_i = \bar{\phi}_i$ (after normalization) can be directly implemented by using any of the two network design methods in subsection III.B.

Analogous conclusions hold for tapering step-sizes. However, choosing the same gain $\varepsilon_n$ for all the agents leads to network centralization through the need for a centralized clock determining $n$. This problem can be overcome by an asynchronous strategy in which the step-size of the $i$-th agent is given by $\varepsilon_i(\Gamma_i(n))$, where $\Gamma_i(n)$ is the number of updates of the agent $i$ up to the instant $n$. General properties of such schemes are discussed in \cite{24}.

In \cite{12} it has been adopted that $\varepsilon_i(\Gamma_i(n)) = 1/\Gamma_i(n)$; it is interesting to note that in this case $|1/\Gamma_i(n) - 1/n| = o(1/n^\lambda)$, where $\lambda > 1$, providing an additional “compensation” for $\bar{d}_i$, and allowing direct comparisons with the exposed design methodology for the constant step-size case.

5 Simulation Results

*Example 1:* A sensor network with $N = 10$ nodes with a Geometric Random Graph topology has been simulated and modified in such a way that a part of two-way connections is randomly transformed into one-way connections. Algorithm AUC has been applied to distributed parameter estimation with $f^i(X_n^i, \xi_n^i) = \mu_n^i - X_n^i$, where $\mu_n^i$ is a random variable $\mu_n^i \sim \mathcal{N}(m_i, \sigma^2_i)$, $m_i$ and $\sigma_i$ being randomly selected in advance from the intervals $[3,7]$ and $[1,5]$, respectively. The network parameters have been chosen in such a way as to achieve equal weights $\bar{\phi}_i \bar{d}_i = \bar{\phi}_j \bar{d}_j$, $i, j = 1, \ldots, N$ (see (8)).

As described above, at the first step, probabilities $p_i$ have been chosen, determining implicitly the values of $\bar{d}_i$, $i = 1, \ldots, N$. At the second step, the asymptotic values $\bar{\varepsilon}_i = \bar{d}_i^{-1} / \sum_{j=1}^{N} \bar{d}_j^{-1}$ have been used to define parameters $\beta_i$, $i = 1, \ldots, N$, according to Case b) in subsection III.B. One degree of freedom in the equation $\dot{\beta} = 0$ has been used to set $\max_i \beta_i$ at its maximal value close to one, in order to maximize in such a way values of all $\beta_i$ and to speed up convergence (giving relatively more weight to all local updates). Fig. 1 gives an illustration of the algorithm behavior; the straight line defines the asymptotic parameter value at consensus $\sum_i \bar{\phi}_i \bar{d}_i m_i$.

Algorithm ACU has the same asymptotic ODE and very similar behavior. However, we have found ACU to be practically slightly superior, having in mind one update per iteration more and explicit noise averaging by making
convexification at the second step (this becomes more visible in networks with high connectedness).

Example 2: The same network, but with fixed $m_i = 5$, $i = 1, \ldots, N$, has been used to check a possibility to optimize network performance using the model (12). Having in mind that $E\{f^i(X^i_n, \xi^i_n)\}|_{X^i_n = X^*} = 0$, the choice of $\phi_i$ and $d_i$ is irrelevant for the parameter value at convergence. In this situation, the exposed method for achieving higher convergence rate starting from (13) has been applied. The optimal values of $\phi_i$ are in this case defined by $\phi^*_i = d^{-2}_i \sigma^{-2}_i / \sum_j d^{-2}_j \sigma^{-2}_j$, $i = 1, \ldots, N$. Fig. 2 shows that such an approximate optimization provides a substantial advantage with respect to the case of equal values of $\phi^*_i$ (dotted lines). In order to make clear the overall performance of the algorithm using the broadcast gossip scheme, the case in which $A_n = \tilde{A}$ and all agents update their estimates for all $n$ is illustrated by solid lines in Fig. 2; the advantage over the gossip scheme is obvious, as expected. In order to show that the weak convergence results can be directly extended to the case of tapering step-size, in Fig. 3 curves analogous to those in Fig. 2 are presented, but with $\varepsilon_n = 1/n$. Convergence to the true parameter value is evident. Curves very similar to those presented in Figs. 2 and 3 have been obtained for the disagreement between the nodes.

6 Conclusion

In this paper distributed stochastic approximation algorithms based on asynchronous broadcast gossip on networks represented by digraphs are considered. Almost sure convergence of the main gossip scheme is proved. Weak convergence results are then derived, resulting in the formulation of the limit ODE depending on the network parameters and probabilities of the nodes to communicate and update their estimates. Two network design schemes are
Figure 2: Mean-square error for AUC, constant step-size: constant consensus matrices (solid lines), gossip (dotted lines); equal weights (A), optimized network parameters (B)

Figure 3: Mean-square error for AUC, tapering step-size $\varepsilon_n = 1/n$: constant consensus matrices (solid lines), gossip (dotted lines); equal weights (A), optimized network parameters (B)
proposed for ensuring the desired asymptotic behavior of the gossip scheme itself, on one side, and of the parameter estimates, on the other. Using a limit stochastic differential equation for the normalized asymptotic error, a method for convergence rate improvement is also proposed.

Further efforts could be oriented towards a more detailed treatment of the case of tapering step-sizes defined asynchronously.

References

[1] H. F. Chen, *Stochastic Approximation and its Applications*. Dordrecht, the Netherlands: Kluwer Academic, 2002.

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

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

[4] J. N. Tsitsiklis, “Problems in decentralized decision making and computation,” Ph.D. dissertation, Dep. Electrical Eng. Comput. Sci., M.I.T., Cambridge, MA, 1984.

[5] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Prentice Hall, 1989.

[6] J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans, “Distributed asynchronous deterministic and stochastic gradient optimization algorithms,” *IEEE Trans. Autom. Control*, vol. 31, pp. 803–812, 1986.

[7] 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.

[8] ——, “Stochastic approximation algorithms for parallel and distributed processing,” *Stochastics*, vol. 22, pp. 219–250, 1987.

[9] A. Nedic, A. Ozdaglar, and P. Parrilo, “Constrained consensus and optimization in multi-agent networks,” *IEEE Trans. on Autom. Contr.*, vol. 55, pp. 922–938, 2010.

[10] A. Nedic and A. Ozdaglar, “Distributed subgradient methods for multi-agent optimization,” *IEEE Trans. on Autom. Contr.*, vol. 54, pp. 48–61, 2009.

[11] S. Ram, A. Nedic, and V. Veeravalli, “Distributed stochastic subgradient projection algorithms for convex optimization,” *Journal of Optim. Theory and Appl.*, vol. 147, pp. 516–545, 2010.
[12] A. Nedic, “Asynchronous broadcast-based convex optimization over a network,” *IEEE Trans. on Autom. Contr.*, vol. 56, pp. 1337–1351, 2011.

[13] V. D. Blondel, J. M. Hendrickx, A. Olshevsky, and J. N. Tsitsiklis, “Convergence in multiagent coordination, consensus and flocking,” in *Proc. IEEE Conf. Decision and Control*, 2005.

[14] P. Bianchi, G. Fort, and W. Hachem, “Performance of a distributed stochastic approximation algorithm,” *IEEE Trans. on Inf. Theory*, vol. 59, pp. 7405–7418, 2013.

[15] P. Bianchi and J. Jakubowicz, “On the convergence of a multi-agent projected stochastic gradient algorithm for nonconvex optimization,” *IEEE Trans. on Automatic Control*, vol. 58, pp. 391–405, 2013.

[16] S. S. Stanković, M. S. Stanković, and D. M. Stipanović, “Decentralized parameter estimation by consensus based stochastic approximation,” *IEEE Trans. Autom. Control*, vol. 47, pp. 531–543, 2011.

[17] S. Tu and A. Sayed, “Diffusion strategies outperform consensus strategies for distributed estimation over adaptive networks,” *IEEE Trans. on Sign. Proc.*, vol. 60, pp. 6217–6233, 2012.

[18] I. Lobel and A. Ozdalgar, “Distributed subgradient methods for convex optimization over random networks,” *IEEE Trans. Automat. Contr.*, vol. 56, pp. 1291–1306, 2011.

[19] T. C. Aysal, M. E. Yildriz, A. D. Sarwate, and A. Scaglione, “Broadcast gossip algorithms for consensus,” *IEEE Trans. Signal Process.*, vol. 57, pp. 2748–2761, 2009.

[20] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, “Randomized gossip algorithms,” *IEEE Trans. Inf. Theory*, vol. 52, pp. 2508–2530, 2006.

[21] S. S. Stanković, M. S. Stanković, and D. M. Stipanović, “Consensus based overlapping decentralized estimation with missing observations and communication faults,” *Automatica*, vol. 45, pp. 1397–1406, 2009.

[22] Z. Lin, “Coupled dynamic systems: from structure towards stability and stabilizability,” Ph.D. dissertation, University of Toronto, 2006.

[23] A. Berman and R. J. Plemmons, *Nonnegative Matrices in Mathematical Sciences*. New York: Academic Press, 1979.

[24] V. Borkar, “Asynchronous stochastic approximation,” *SIAM Journal on Control And Optimization*, vol. 36, pp. 840–851, 1998.