Distributed Subgradient Methods for Convex Optimization Over Random Networks

Ilan Lobel and Asuman Ozdaglar, *Member, IEEE*

***Abstract*—We consider the problem of cooperatively minimizing the sum of convex functions, where the functions represent local objective functions of the agents. We assume that each agent has information about his local function, and communicate with the other agents over a time-varying network topology. For this problem, we propose a distributed subgradient method that uses averaging algorithms for locally sharing information among the agents. In contrast to previous works on multi-agent optimization that make worst-case assumptions about the connectivity of the agents (such as bounded communication intervals between nodes), we assume that links fail according to a given stochastic process. Under the assumption that the link failures are independent and identically distributed over time (possibly correlated across links), we provide almost sure convergence results for our subgradient algorithm.**

I. Introduction

THERE has been considerable interest in cooperative control problems in large-scale networks. Objectives range from detecting and computing some information using a network of sensors to allocating resources in large communication networks. A common feature of these problems is the need for a solution method that is completely decentralized and is not computationally heavy, so that simple sensors or busy network servers are not overburdened by it. We shall call these sensors (or servers or routers) our agents, or alternatively, the nodes of the network.

Such large networks are also often *ad hoc* in nature: the availability of a communication link between a given pair of agents is usually random. In the case of sensor networks, the nodes routinely shut down their antennas in order to conserve energy and, even when both sensors are trying to communicate with each other, there are sometimes physical obstructions that block the wireless channel.

These considerations necessitate designing methods that solve optimization problems in a decentralized way using local information and taking into consideration the fact that communication link between agents in the network is not always available.

Manuscript received January 22, 2009; revised December 04, 2009; accepted July 27, 2010. Date of publication November 09, 2010; date of current version June 08, 2011. This work was supported in part by the National Science Foundation under Career grant DMI-0545910, in part by the DARPA ITMANET program, and in part by Microsoft Research New England Lab. Recommended by Associate Editor I. Paschalidis.

I. Lobel is with the Information, Operations and Management Sciences Department, Stern School of Business, New York University, New York, NY 10012 (e-mail: ilobel@stern.nyu.edu).

A. Ozdaglar is with the Laboratory for Information and Decision Systems, Electrical Engineering and Computer Science Department, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: asuman@mit.edu).

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

Digital Object Identifier 10.1109/TAC.2010.2091295

In this paper, we develop distributed subgradient methods for cooperatively optimizing a global objective function, which is a function of the individual agent objective functions. These methods operate over a network with randomly varying connectivity. Our approach builds on the seminal work by Tsitsiklis [22] (see also Tsitsiklis *et al.* [23], Bertsekas and Tsitsiklis [2]), which developed a general framework for parallel and distributed computation among different processors, and on the recent work by Nedić and Ozdaglar [14], which studied a distributed method for cooperative optimization in multi-agent environments. Both of these works make worst-case assumptions about communication link availability, such as bounded intercommunication intervals between any two neighboring nodes in the network. In contrast, in this paper, we assume that the communication link availability is represented by a stochastic process. As such, the presence of a communication link between any two nodes at a given time period is a random event, which is possibly correlated with the availability of other communication links in the same interval. Our work is also related to the literature on randomized consensus algorithms where the randomness may be due to the choice of the randomized communication protocol (as in the gossip algorithms studied in Boyd *et al.* [5]), or due to the unpredictability in the environment that the information exchange takes place (see Hatano and Mesbahi [9], Wu [25], Tahbaz-Salehi and Jadbabaie [21] and Fagnani and Zampieri [7]). Our paper uses a random graph model which is similar to [21], and presents distributed subgradient methods that can optimize general convex (not necessarily smooth) local objective functions.

More specifically, our model involves a set of agents whose goal is to cooperatively minimize a convex objective function $\sum_{i=1}^{n} f_i(x)$, where n is the number of agents and the function $f_i : \mathbb{R}^m \to \mathbb{R}$ is the local objective of agent i, known only by this agent. Such problems arise in congestion control problems in wireline networks, where heterogeneous users adjust their flow rates to maximize their utility minus latency they experience along their routes (see Kelly *et al.* [11]). Another application area is distributed sensor networks where spatially distributed sensors use their local measurements to estimate certain quantities. The objective function of sensor i can be represented as $f_i(x) = \mathrm{E}[F_i(R_i, x)]$, where R_i is some random process observed locally by agent i and the function $F_i(R_i, x)$ captures the quality of agent i's estimates (see [19]).

Our algorithm works as follows: each agent i maintains a pair of estimates $x_i(k)$ and $\tilde{x}_i(k)$ of the optimal solution of the optimization problem at each point in time $k \geq 0$. Agent i updates the estimate $x_i(k)$ by averaging the value of $x_i(k)$ with the estimates of neighboring nodes in the network and by taking a step in the direction given by the negative of the subgradient of function f_i at value $x_i(k)$. The estimate $\tilde{x}_i(k)$ is a long-run (time) average of the values of $x_i(k)$.

Using a diminishing stepsize, we prove that agent estimates converge to the same point in the optimal solution set with probability one. For a constant stepsize, we show that, with probability 1, the limit superior of the objective function values of agents' (averaged) estimates lies in a neighborhood of the optimal value of the problem. We also characterize explicitly the error neighborhood in terms of the parameters of the problem.

Our work is related to the literature on reaching consensus on a particular scalar value or computing exact averages of the initial values of the agents, which has attracted much recent attention as natural models of cooperative behavior in networked-systems (see Vicsek *et al.* [24], Jadbabaie *et al.* [10], Olfati-Saber and Murray [16], Cao *et al.* [6], Olshevsky and Tsitsiklis [17], [18], and Nedić *et al.* [13]). Our work is also related to the *utility maximization* framework for resource allocation in networks (see Kelly *et al.* [11], Low and Lapsley [12], Srikant [20], and Chiang *et al.* [8]). In contrast to this literature, we consider a model with general (convex) agent performance measures.

The remainder of this paper is organized as follows: In Section II, we formally introduce the model. Sections III–V build the tools that we use to analyze our model: Section III develops some results on the communication networks, Section IV establishes some preliminary results about products of random matrices and Section V studies the convergence properties of the iterates of the subgradient method. Section VI concludes the paper.

1) Basic Notation and Notions: A vector is viewed as a column vector, unless clearly stated otherwise. We denote by x^i or $[x]^i$ the i-th component of a vector x. When $x^i \geq 0$ for all components i of a vector x, we write $x \geq 0$. For a matrix A, we write A_{ij} or $[A]_i^j$ to denote the matrix entry in the i-th row and j-th column. For an ordered pair $e = (i, j)$, we also use the notation A_e to denote the (i, j) entry of matrix A. We write $[A]_i$ to denote the i-th row of the matrix A, and $[A]^j$ to denote the j-th column of A.

We denote the nonnegative orthant by $\mathbb{R}_+^m$, i.e., $\mathbb{R}_+^m = \{x \in \mathbb{R}^m \mid x \geq 0\}$. We write x' to denote the transpose of a vector x. The scalar product of two vectors $x, y \in \mathbb{R}^m$ is denoted by $x'y$. We use $\|x\|$ to denote the standard Euclidean norm, $\|x\| = \sqrt{x'x}$. We write $\|x\|_\infty$ to denote the max norm, $\|x\|_\infty = \max_{1\leq i\leq m} |x_i|$.

A vector $a \in \mathbb{R}^n$ is said to be a *stochastic vector* when its components a_i, $i = 1, \ldots, n$, are nonnegative and their sum is equal to 1, i.e., $\sum_{i=1}^n a_i = 1$. A square $n \times n$ matrix A is said to be a *stochastic matrix* when each row of A is a stochastic vector. A square $m \times m$ matrix A is said to be a *doubly stochastic* matrix when both A and A' are stochastic matrices.

For a function $F : \mathbb{R}^m \to (-\infty, \infty]$, we denote the domain of F by $\mathrm{dom}(F)$, where

$$\mathrm{dom}(F) = \{x \in \mathbb{R}^m \mid F(x) < \infty\}.$$

We use the notion of a subgradient of a *convex* function $F(x)$ at a given vector $\bar{x} \in \mathrm{dom}(F)$. We say that $s_F(\bar{x}) \in \mathbb{R}^n$ *is a subgradient of the function* F at $\bar{x} \in \mathrm{dom}(F)$ when the following relation holds:

$$F(\bar{x}) + s_F(\bar{x})'(x - \bar{x}) \leq F(x) \qquad \text{for all } x \in \mathrm{dom}(F). \quad (1)$$

The set of all subgradients of F at $\bar{x}$ is denoted by $\partial F(\bar{x})$ (see [1]).

II. The Model

We consider a network with a set of nodes (or agents) $\mathcal{N} = \{1, \ldots, n\}$. The goal of agents is to collectively minimize a common additive cost. Each agent has information only about one cost component, and minimizes that component while exchanging information with other agents. In particular, the agents want to solve the following unconstrained optimization problem:

$$\begin{aligned} &\text{minimize} \quad \sum_{i=1}^n f_i(x) \\ &\text{subject to} \quad x \in \mathbb{R}^m \end{aligned} \qquad (2)$$

where each $f_i : \mathbb{R}^m \to \mathbb{R}$ is a convex function. We denote the optimal value of this problem by f^*, which we assume to be *finite*. We also denote the optimal solution set by X^*, i.e., $X^* = \{x \in \mathbb{R}^m \mid \sum_{i=1}^n f_i(x) = f^*\}$. Throughout the paper, we assume that the optimal solution set X^* is *nonempty*.

Each agent i starts with some initial estimate (or information) about the optimal solution of problem (2), which we denote by $x_i(0) \in \mathbb{R}^m$. Agents communicate with neighboring agents and update their estimates at discrete instances $t_0, t_1, t_2, \ldots$. We discretize time according to these instances and denote the estimate of agent i at time t_k as $x_i(k)$.

At each time $k+1$, we assume that agent i receives information $x_j(k)$ from neigboring agents j and updates his estimate. We represent this update rule as

$$x_i(k+1) = \sum_{j=1}^n a_{ij}(k) x_j(k) - \alpha(k) d_i(k) \qquad (3)$$

where the vector $(a_{i1}(k), \ldots, a_{\mathrm{in}}(k))'$ is a vector of weights for agent i and the sequence $\{\alpha(k)\}$ establishes the stepsizes. The vector $d_i(k)$ is a subgradient of agent i objective function $f_i(x)$ at his current estimate $x = x_i(k)$. This update rule represents a combination of new information from other agents in the network and an optimization step along the subgradient of the local objective function of agent i. We note that the widely studied linear averaging algorithms for consensus (or agreement) problems are special cases of the *optimization update rule* (3) when the functions f_i are identically equal to zero; see Jadbabaie *et al.* [10] and Blondel *et al.* [4].

Let $x^l(k)$ denote the vector comprised of the lth components of all agent estimates at time k, i.e., $x^l(k) = (x_1^l(k), \ldots, x_n^l(k))$ for all $l = 1, \ldots, m$. The update rule in (3) implies that the component vectors of agent estimates evolve according to

$$x^l(k+1) = A(k) x^l(k) - \alpha(k) d^l(k)$$

where the vector $d^l(k) = \left(d_1^l(k), \ldots, d_n^l(k)\right)$ is a vector of the lth component of the subgradient vector of each agent, and the matrix $A(k)$ is a matrix with components $A(k) = [a_{ij}(k)]_{i,j\in\mathcal{N}}$.

We adopt a probabilistic approach to model the availability of communication links between different agents. In particular, we assume that the matrix $A(k)$ is a *random matrix* that describes the time-varying connectivity of the network. The Section III describes our assumptions on the random weight matrices $A(k)$.

A. Model of Communication

Assumption 1: (Weights) Let $\mathcal{F} = (\Omega, \mathcal{B}, \mu)$ be a probability space such that Ω is the set of all $n \times n$ stochastic matrices, $\mathcal{B}$ is the Borel σ-algebra on Ω and μ is a probability measure on $\mathcal{B}$.

(a) There exists a scalar γ with $0 < \gamma < 1$ such that $A_{ii} \geq \gamma$ for all i with probability 1.

(b) For all $k \geq 0$, the matrix $A(k)$ is drawn independently from probability space $\mathcal{F}$.

The assumption that $A(k)$ is drawn from the set Ω of stochastic matrices implies that each agent takes a convex combination of the information he receives from his neighbors in the update rule (3). Assumption 1(a) ensures that each agent gives significant weight to his own estimate $x_i(k)$ at each time k. Assumption 1(b) states that the induced graph, i.e., the graph $(\mathcal{N}, \mathcal{E}_+(k))$ where $\mathcal{E}_+(k) = \{(j,i) \mid a_{ij}(k) > 0\}$, is a random graph that is independent and identically distributed over time k. Note that this assumption allows the edges of the graph $(\mathcal{N}, \mathcal{E}_+(k))$ at any time k to be correlated [see also Hatano and Mesbahi [9] for a more specialized random graph model, where each edge is realized randomly and independently of all other edges in the graph $(\mathcal{N}, \mathcal{E}_+(k))$ (i.e., according to an Erdös-Rényi random graph model), and Wu [25] and Tahbaz-Salehi and Jadbabaie [21] for similar random graph models]. Formally, we define a product probability space $(\Omega^\infty, \mathcal{B}^\infty, \mu^\infty) = \prod_{k=0}^{\infty}(\Omega, \mathcal{B}, \mu)$. Assumption 1(b) implies that the entire sequence $\{A(k)\}$ is drawn from this product probability space. We denote a realization in this probability space by $A^\infty = \{A(k)\} \in \Omega^\infty$.

We next describe our connectivity assumption among the agents. To state this assumption, we consider the expected value of the random matrices $A(k)$, which in view of the independence assumption over k, can be represented as

$$\tilde{A} = E[A(k)] \qquad \text{for all} \quad k \geq 0. \tag{4}$$

We consider the edge set induced by the positive elements of the matrix $\tilde{A}$, i.e.,

$$\tilde{\mathcal{E}} = \{(j,i) \mid \tilde{A}_{ij} > 0\},$$

and the corresponding graph $(\mathcal{N}, \tilde{\mathcal{E}})$, which we refer to as the *mean connectivity graph*.

Assumption 2: (Connectivity) The mean connectivity graph $(\mathcal{N}, \tilde{\mathcal{E}})$ is strongly connected.

This assumption imposes a mild connectivity condition among the agents and ensures that in expectation, the information of an agent i reaches every other agent i directly or indirectly through a directed path.

Finally, we assume without loss of generality that the scalar $\gamma > 0$ of part (a) of the Weights Assumption [cf. Assumption 1(a)] provides a uniform lower bound on the positive elements of the matrix $\tilde{A}$, i.e.,

$$\min_{(j,i)\in\tilde{\mathcal{E}}} \frac{\tilde{A}_{ij}}{2} \geq \gamma. \tag{5}$$

III. Network Communication Preliminaries

This section constructs random communication events that have the following property: if one such event occurs, then information has propagated from each agent to every other agent. We establish bounds on the probability of such an event occurring and the 'amount' of information that propagates when it happens. These events are used in forthcoming sections to analyze the convergence of the distributed subgradient method.

We introduce the *transition matrices* $\Phi(k,s)$ for any s and k with $k \geq s \geq 0$ as

$$\Phi(k,s) = A(s)A(s+1)\cdots A(k-1)A(k) \tag{6}$$

for all s and k with $k \geq s$, where $\Phi(k,k) = A(k)$ for all k. Using the transition matrices, we can relate the generated estimates of (3) as follows: for any $i \in \mathcal{N}$, and any s and k with $k \geq s \geq 0$

$$x_i(k+1) = \sum_{j=1}^{n}[\Phi(k,s)]_{ij}x_j(s) - \sum_{r=s+1}^{k}\sum_{j=1}^{n}[\Phi(k,r)]_{ij}\alpha(r-1)d_j(r-1) - \alpha(k)d_i(k) \tag{7}$$

(see [14] for more details). As seen from the preceding relation, we need to understand the convergence properties of the transition matrices $\Phi(k,s)$ to study the asymptotic behavior of the estimates $x_i(k)$. These properties are established in the following two lemmas. Deterministic variations of these lemmas have been proven in [14].

The first lemma provides positive lower bounds on each entry (i,j) of the transition matrix $\Phi(k,s)$. Such bounds are obtained under the condition that the matrix entry $[A(r)]_{ij}$ satisfies $[A(r)]_{ij} \geq \gamma$, for some time r with $s \leq r \leq k$, or equivalently information is exchanged on link (j,i) at time r. We say that link (j,i) is *activated at time* k when $[A(k)]_{ij} \geq \gamma$ and use the edge set $\mathcal{E}(k)$ to identify such edges, i.e., for any $k \geq 0$, the set $\mathcal{E}(k)$ denotes the set of edges induced by the *sufficiently positive* elements of the matrix $A(k)$

$$\mathcal{E}(k) = \{(j,i) \mid [A(k)]_{ij} \geq \gamma\}. \tag{8}$$

Lemma 1: Let Weights Assumption hold [cf. Assumption 1]. The following statements hold with probability one:

(a) $[\Phi(k,s)]_{ii} \geq \gamma^{k-s+1}$ for all i, and s and k with $k \geq s \geq 0$.

(b) $[\Phi(k,s)]_{ij} \geq \gamma^{k-s+1}$ for all s and k with $k \geq s \geq 0$ and all $(j,i) \in \mathcal{E}(r)$ for some $s \leq r \leq k$.

(c) Let $(j,v) \in \mathcal{E}(s)$ for some $s \geq 0$ and $(v,i) \in \mathcal{E}(r)$ for some $r > s$. Then, $[\Phi(k,s)]_{ij} \geq \gamma^{k-s+1}$ for all $k \geq r$.

Proof: For parts (a) and (b), we let s be arbitrary and prove the relations by induction on k.

(a) By the definition of the transition matrices (6) and Assumption 1(a), we have $[\Phi(s,s)]_{ii} = [A(s)]_{ii} \geq \gamma$. Thus, the relation holds for $k = s$.
Now, assume that for some k with $k > s$ we have $[\Phi(k,s)]_{ii} \geq \gamma^{k-s+1}$, and consider $[\Phi(k+1,s)]_{ii}$. By the definition of the matrix $\Phi(k,s)$, we have

$$\begin{aligned}[\Phi(k+1,s)]_{ii} &= \sum_{h=1}^{n}[A(k+1)]_{ih}[\Phi(k,s)]_{hi} \\ &\geq [A(k+1)]_{ii}[\Phi(k,s)]_{ii},\end{aligned}$$

where the inequality follows from the nonnegativity of the entries of $\Phi(k,s)$. By using the inductive hypothesis and the relation $[A(k+1)]_{ii} \geq \gamma$ [cf. Assumption 1(a)], we obtain

$$[\Phi(k+1,s)]_{ii} \geq \gamma^{k-s+2}$$

establishing the relation.

(b) Let $(j,i) \in \mathcal{E}(s)$. Then, by the definition of the set $\mathcal{E}(s)$ and the transition matrices (i.e., $\Phi(s,s) = A(s)$), it follows that the relation $[\Phi(k,s)]_{ij} \geq \gamma^{k-s+1}$ holds for $k = s$ and any $(j,i) \in \mathcal{E}(s)$. Assume now that for some $k > s$ and all $(j,i) \in \mathcal{E}(r)$ with $s \leq r \leq k$, we have $[\Phi(k,s)]_{ij} \geq \gamma^{k-s+1}$. Consider $k+1$, and let $(j,i) \in \mathcal{E}(r)$ for some $s \leq r \leq k+1$. There are two possibilities: $s \leq r \leq k$ or $r = k+1$.
Suppose first that $s \leq r \leq k$. Then by the induction hypothesis, we have $[\Phi(k,s)]_{ij} \geq \gamma^{k-s+1}$. Therefore

$$\begin{aligned}[\Phi(k+1,s)]_{ij} &= \sum_{h=1}^{n}[A(k+1)]_{ih}[\Phi(k,s)]_{hj} \\ &\geq [A(k+1)]_{ii}[\Phi(k,s)]_{ij} \geq \gamma^{k-s+2},\end{aligned}$$

where the second inequality follows from the fact that $[A(k+1)]_{ii} \geq \gamma$ [cf. Assumption 1(a)].
Suppose now that $r = k+1$, i.e., $(j,i) \in \mathcal{E}(k+1)$. By the definition of $\mathcal{E}(k+1)$, we have $[A(k+1)]_{ij} \geq \gamma$. Moreover, since $[\Phi(k,s)]_{jj} \geq \gamma^{k-s+1}$ by part (a) of the lemma, we obtain

$$\begin{aligned}[\Phi(k+1,s)]_{ij} &= \sum_{h=1}^{n}[A(k+1)]_{ih}[\Phi(k,s)]_{hj} \\ &\geq [A(k+1)]_{ij}[\Phi(k,s)]_{jj} \geq \gamma^{k-s+2}\end{aligned}$$

completing the induction.

(c) Let $(j,v) \in \mathcal{E}(s)$ for some $s \geq 0$ and $(v,i) \in \mathcal{E}(r)$ for some $r > s$. We have

$$\begin{aligned}[\Phi(k,s)]_{ij} &= \sum_{h=1}^{n}[\Phi(k,s+1)]_{ih}[A(s)]_{hj} \\ &\geq [\Phi(k,s+1)]_{iv}[A(s)]_{vj}.\end{aligned}$$

By the definition of the edge set $\mathcal{E}(s)$, we have $[A(s)]_{vj} \geq \gamma$. By part (b), since $(v,i) \in \mathcal{E}(r)$ and $s < r \leq k$, we have

$$[\Phi(k,s+1)]_{iv} \geq \gamma^{k-s}.$$

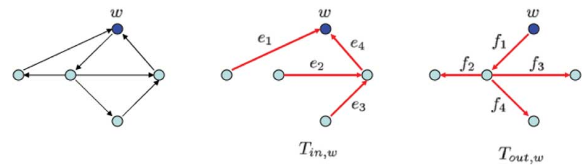

Fig. 1. Strongly connected mean connectivity graph and the two directed spanning trees rooted at node w on this graph. The figure illustrates the labeling of the edges on the in-tree $T_{\mathrm{in},w}$ and the out-tree $T_{\mathrm{out},w}$ according to the procedure described in the text. Note that the edges on all directed paths are labeled in nondecreasing order.

Combining these relations, we obtain

$$[\Phi(k,s)]_{ij} \geq \gamma^{k-s+1}.$$

■

We next construct a probabilistic event in which the edges of the graphs $\mathcal{E}(k)$ are activated over time k in such a way that information propagates from every agent to every other agent in the network.

To define this event, we fix a node $w \in \mathcal{N}$ and consider *two directed spanning trees* in the mean connectivity graph $(\mathcal{N}, \tilde{\mathcal{E}})$: an in-tree rooted at w, denoted by $T_{\mathrm{in},w}$ (i.e., there exists a directed path from every node $i \neq w$ to w on the tree), and an out-tree rooted at w, denoted by $T_{\mathrm{out},w}$ (i.e., there exists a directed path from w to every node $i \neq w$ on the tree). Under the assumption that the mean connectivity graph $(\mathcal{N}, \tilde{\mathcal{E}})$ is strongly connected (cf. Assumption 2), these spanning trees exist and each contain $n-1$ edges (see [3]).

We consider a specific ordering of the edges of these spanning trees. In particular, for the in-tree $T_{\mathrm{in},w}$, we pick an arbitrary leaf node and label the adjacent edge as e_1; then we pick another leaf node and label the adjacent edge as e_2; we repeat this until all leaves are picked. We then delete the leaf nodes and the adjacent edges from the spanning tree $T_{\mathrm{in},r}$, and repeat the same process for the new tree. This edge labeling ensures that on any directed path from a node $i \neq w$ to node w, edges are labeled in nondecreasing order.

Similarly, for the out-tree $T_{\mathrm{out},w}$, we pick a directed path from node w to an arbitrary leaf and sequentially label the edges on the directed path; we then consider a directed path from node w to another leaf and label the unlabeled edges on the path sequentially from the root node w to the leaf;[1] we continue until all directed paths to all the leaves are exhausted. We represent the edges of the two spanning trees with the order described above as

$$\begin{aligned}T_{\mathrm{in},w} &= \{e_1, e_2, \ldots, e_{n-1}\} \\ T_{\mathrm{out},w} &= \{f_1, f_2, \ldots, f_{n-1}\}\end{aligned} \tag{9}$$

(see Fig. 1).

We define the probabilistic event that ensures information exchange across the network as follows. Recall that for any edge $e = (j,i)$, the notation A_e denotes the (i,j) entry of the matrix

[1]Note that this edge labeling ensures that all edges are labeled in a nondecreasing order on this path; otherwise there would exist an "out-of-order" edge on this path, implying that it was labeled before the edges that precede it on the path, i.e., it belongs to another directed path that originates from root node w on the tree $T_{\mathrm{out},w}$, but it can be seen that this creates a cycle on the tree $T_{\mathrm{out},w}$—a contradiction.

A. Given any time $s \geq 0$, we define the following events for all $l = 1, \ldots, n-1$

$$C_l(s) = \{A^\infty \in \Omega^\infty \mid A_{e_l}(s+l-1) \geq \gamma\} \quad (10)$$
$$D_l(s) = \{A^\infty \in \Omega^\infty \mid A_{f_l}(s+(n-1)+l-1) \geq \gamma\} \quad (11)$$

as well as

$$G(s) = \bigcap_{l=1,\ldots,n-1} (C_l(s) \cap D_l(s)) \quad (12)$$

for $s \geq 0$. For all $l = 1, \ldots, n-1$, the event $C_l(s)$ denotes the event that edge $e_l \in T_{\text{in},w}$ is activated at time $s+l-1$, and the event $D_l(s)$ denotes the event that edge $f_l \in T_{\text{out},w}$ is activated at time $s+(n-1)+l-1$. Hence, for any $s \geq 0$, the event $G(s)$ denotes the event in which each edge in the spanning trees $T_{\text{in},w}$ and $T_{\text{out},w}$ are activated sequentially following time s in the order given in (9).

Lemma 2: Let Weights and Connectivity Assumptions hold [cf. Assumptions 1 and 2]. For any $s \geq 0$, let $A^\infty \in G(s)$, where the event $G(s)$ is defined in (12). Then, we have

$$[\Phi(k,s)]_{ij} \geq \gamma^{k-s+1} \text{ for all } i,j, \text{ and } k \geq s+2(n-1)-1.$$

Proof: Let i, j be arbitrary. If $j = i$, then by Lemma 1(a), we have

$$[\Phi(k,s)]_{ii} \geq \gamma^{k-s+1}.$$

Suppose now that $j \neq i$. By Connectivity assumption (cf. Assumption 2), there exists a path $j = j_0 \rightarrow j_1 \rightarrow \cdots j_{k_{\text{in}}-1} \rightarrow j_{k_{\text{in}}} = w$ from $j \rightarrow w$ with edges on the in-tree $T_{\text{in},w}$, i.e., for each edge (j_h, j_{h+1}), $h = 0, \ldots, k_{\text{in}}-1$, there exists some $l(h) = 1, \ldots, n-1$ such that $(j_h, j_{h+1}) = e_{l(h)}$ [cf. (9)]. Moreover, in view of the ordering of the edges on the in-tree $T_{\text{in},w}$, it follows that the sequence $\{l(h)\}_{h=0,\ldots,k_{\text{in}}}$ is nondecreasing. Since by assumption $A^\infty \in G(s)$, it follows from the definition of the event $G(s)$ [cf. (12)] that

$$A_{(j_{h+1},j_h)}(s+l(h)-1) \geq \gamma \quad \text{for all } h = 0, \ldots, k_{\text{in}}$$

and for some nondecreasing sequence $\{l(h)\}$ that belongs to the set $\{1, \ldots, n-1\}$. By the definition of the edge set $\mathcal{E}(k)$, i.e.,

$$\mathcal{E}(k) = \{(j,i) \mid [A(k)]_{ij} \geq \gamma\}$$

[cf. (8)], this implies that

$$(j_h, j_{h+1}) \in \mathcal{E}(s+l(h)-1) \qquad \text{for all } h = 0, \ldots, k_{\text{in}} \quad (13)$$

for some nondecreasing sequence $\{l(h)\} \subset \{1, \ldots, n-1\}$.

Similarly, by the connectivity assumption (cf. Assumption 2), there exists a path $w = i_0 \rightarrow i_1 \rightarrow \cdots i_{k_{\text{out}}-1} \rightarrow i_{k_{\text{out}}} = i$ from $w \rightarrow i$ with edges on the out-tree $T_{\text{out},w}$, i.e., for each edge (i_h, i_{h+1}), $h = 0, \ldots, k_{\text{out}}-1$, there exists some $\bar{l}(h) = 1, \ldots, n-1$ such that $(i_h, i_{h+1}) = f_{\bar{l}(h)}$ [cf. (9)]. The ordering of the edges on the out-tree $T_{\text{out},w}$ implies that the sequence $\{\bar{l}(h)\}_{h=0,\ldots,k_{\text{out}}}$ is nondecreasing. Using again the assumption $A^\infty \in G(s)$ and the definition of the event $G(s)$ [cf. (12)], we have

$$A_{(i_{h+1},i_h)}(s+(n-1)+\bar{l}(h)-1) \geq \gamma \text{ for } h = 0, \ldots, k_{\text{out}}$$

from which we obtain

$$(i_h, i_{h+1}) \in \mathcal{E}(s+(n-1)+\bar{l}(h)-1) \quad \text{for all } h = 0, \ldots, k_{\text{out}} \quad (14)$$

for some nondecreasing sequence $\{\bar{l}(h)\} \subset \{1, \ldots, n-1\}$.

Combining (13)–(14) with Lemma 1(c), it follows that, for all $k \geq s+2(n-1)-1$

$$[\Phi(k,s)]_{ij} \geq \gamma^{k-s+1} \quad \text{for all } i,j$$

establishing the desired relation. ■

The previous lemma states that for any $s \geq 0$, if the event $G(s)$ occurs, then every entry of the transition matrix $\Phi(k,s)$ is uniformly bounded away from 0 for sufficiently large k. In the next lemma, we show that the event $G(s)$ occurs with positive probability and provide a positive uniform lower bound on the probability over all s.

Lemma 3: Let Weights and Connectivity Assumptions hold [cf. Assumptions 1 and 2]. For any $s \geq 0$, the following hold:

(a) For all $l = 1, \ldots, n-1$, The events $C_l(s)$ and $D_l(s)$ are mutually independent and

$$P(C_l(s)) \geq \gamma, \text{ and } P(D_l(s)) \geq \gamma.$$

(b) $P(G(s)) \geq \gamma^{2(n-1)}$.

Proof:

(a) Given any $s \geq 0$, each event $C_l(s)$ and $D_l(s)$, for $l = 1, \ldots, n-1$, is associated with a distinct time, i.e., each such event belongs to the σ-algebra generated by $A(s)$. By Assumption 1(b), it follows that the events $C_l(s)$ and $D_l(s)$ for all $l = 1, \ldots, n-1$ are mutually independent. We next establish the lower bound on $P(C_l(s))$. By the definition of the event $C_l(s)$ [cf. (10)], we have for any $s \geq 0$ and $l = 1, \ldots, n-1$

$$\begin{aligned} P(C_l(s)) &= P(A_{e_l}(s+l-1) \geq \gamma) \\ &= P(1 - A_{e_l}(s+l-1) \leq 1-\gamma) \\ &\geq P(1 - A_{e_l}(s+l-1) < 1-\gamma) \\ &= 1 - P(1 - A_{e_l}(s+l-1) \geq 1-\gamma). \end{aligned}$$

The Markov inequality states that for any nonnegative random variable Y with a finite mean $E[Y]$, the probability that the outcome of the random variable Y exceeds any given scalar $\delta > 0$ satisfies

$$P(\{Y \geq \delta\}) \leq \frac{E[Y]}{\delta}.$$

By applying the Markov inequality to the random variable $1 - A_{e_l}(s+l-1)$ (which is nonnegative and has finite expectation in view of the assumption that the matrix $A(k)$ is a stochastic matrix for all k), we obtain

$$P(1 - A_{e_l}(s+l-1) \geq 1-\gamma) \leq \frac{1 - E[A_{e_l}(s+l-1)]}{1-\gamma}.$$

Combining the preceding two relations, we have

$$P(C_l(s)) \geq 1 - \frac{1 - E[A_{e_l}(s+l-1)]}{1-\gamma}.$$

By the independence of the matrices $A(k)$ and the definition of the mean matrix $\tilde{A}$ [cf. (4)], we have

$$E[A_{e_l}(s+l-1)] = [\tilde{A}]_{e_l} \geq 2\gamma$$

where the inequality follows from the bound in (5). Using this bound in the preceding relation, we obtain

$$P(C_l(s)) \geq 1 - \frac{1-2\gamma}{1-\gamma} = \frac{\gamma}{1-\gamma} \geq \gamma$$

where the last inequality follows from the fact that $0 < \gamma < 1$ [cf. Assumption 1(a)].

(b) By the definition of the event $G(s)$ in (12), and the independence of the events $C_l(s)$ and $D_l(s)$, we immediately have

$$P(G(s)) = \prod_{l=1,\ldots,n-1} P(C_l(s))\, P(D_l(s)) \geq \gamma^{2(n-1)}$$

where the second inequality follows from part (a) of this lemma. ■

Thus, we have constructed an event $G(s)$ for each $s \geq 0$ such that $P(G(s)) \geq \gamma^{2(n-1)}$ and, if it occurs, it implies that information is exchanged between all agents, i.e.,

$$[\Phi(s+2(n-1)-1, s)]_{ij} \geq \gamma^{2(n-1)} \quad \text{for } (i,j) \in \{1,\ldots,n\}^2.$$

IV. Random Matrices

In this section, we analyze some properties of products of random matrices that are essential to our analysis. We start by analyzing sequences of deterministic matrices and then proceed to use large deviations theory to analyze sequences of random matrices.

The following lemma is based on a similar result from Nedic and Ozdaglar [14] and relates to a seminal result from Tsitsiklis [22]. We skip the proof because it is very similar to the proof of Lemma 3 in [14].

Lemma 4: Let $\{D_k\}$ be a sequence of stochastic matrices (with n rows and columns) and let $\delta > 0$ be a scalar. Assume that for any $k \geq 0$ and any element $(i,j) \in \{1,\ldots,n\}^2$, $[D_k]_i^j \geq \delta$. Then

(a) The limit $\bar{D} = \lim_{k\to\infty} D_k \cdots D_1$ exists.

(b) The matrix $\bar{D}$ is stochastic and its rows are identical.

(c) The convergence of $D_k \cdots D_1$ to $\bar{D}$ is geometric:

$$\max_{(i,j)\in\{1,\ldots,n\}^2} \left| [D_k \cdots D_1]_i^j - [\bar{D}]_i^j \right| \leq 2\left(1 + \frac{1}{\delta}\right)(1-\delta)^k$$

for all $k \geq 1$.

To obtain convergence of the subgradient method, we need the matrices $\{A(k)\}$ to be doubly stochastic.

Assumption 3: (Doubly Stochastic Weights) Let the weight matrices $A(k)$, $k = 0, 1, \ldots$ satisfy Weights Rule [cf. Assumption 1]. Assume further that the matrices $A(k)$ are doubly stochastic with probability 1.

One sufficient condition for a stochastic matrix to be doubly stochastic is symmetry. If every pair of agents coordinate their weights when they communicate so that they use the same coefficients, i.e., for each $k \geq 0$, $a_{ij}(k) = a_{ji}(k)$ for all $(i,j) \in \{1,\ldots,n\}^2$ with probability 1, then doubly stochasticity is satisfied.[2]

Lemma 5: $\{D_k\}$ be a sequence of doubly stochastic matrices (with n rows and columns) such that the product $D_k \cdots D_1$ converges to $\bar{D}$. Then, any element $(i,j) \in \{1,\ldots,n\}^2$ of $\bar{D}$ satisfies $[\bar{D}]_i^j = 1/n$. Furthermore, if for all k, all elements of D_k are greater than or equal to some $\delta > 0$, i.e., $[D_k]_i^j \geq \delta$ for all $i, j \in \mathcal{N}$, then for all $k \geq 1$

$$\max_{(i,j)\in\{1,\ldots,n\}^2} \left| [D_k \cdots D_1]_i^j - \frac{1}{n} \right| \leq 2\left(1 + \frac{1}{\delta}\right)(1-\delta)^k.$$

Proof: Since the matrix D_k is doubly stochastic for all k, the limit matrix $\bar{D}$ is also doubly stochastic. In view of Lemma 4(b), the limit matrix $\bar{D}$ has identical rows, i.e., there exists a vector ϕ such that $\bar{D} = \phi e'$. Therefore, we have $\phi e' e = e$, implying that $\phi = (1/n)e$. The second claim of the lemma follows immediately from $[\bar{D}]_i^j = 1/n$ for all $i, j \in \mathcal{N}$ and Lemma 4(c). ■

Lemma 5 suggests a way to measure how distant a product of doubly matrices is from its limit. Let us then introduce the metric

$$b(k,s) = \max_{(i,j)\in\{1,\ldots,n\}^2} \left| [\Phi(k,s)]_i^j - \frac{1}{n} \right| \quad \text{for all } k \geq s. \tag{15}$$

The following lemma states that if t independent events of the form $G(s_i)$, for $i = 1,\ldots,t$, occur between times r and k with $k > r \geq 0$, then $b(k,r)$ decays geometrically in t. This is a lemma about deterministic matrices, because the result is conditional on the occurrence of the random events $G(s_1)$, $i = 1,\ldots,t$.

Lemma 6: Let Connectivity and Doubly Stochastic Weights Assumptions hold [cf. Assumptions 2 and 3]. Let t be a positive integer and consider scalars $r < s_1 < s_2 < \cdots < s_t < k$. Further assume that $s_i + 2(n-1) \leq s_{i+1}$ for each $i = 1,\ldots,t-1$ and $s_t \leq k$. For a fixed realization A^∞, let $b(k,r)$ be defined as in (15) and assume that events $G(s_i)$ occur for each $i \in 1,\ldots,t$. Then

$$b(k,r) \leq 2\left(1 + \frac{1}{\gamma^{2(n-1)}}\right)(1-\gamma^{2(n-1)})^t. \tag{16}$$

Proof: For the fixed realization A^∞, define the following t matrices:

$$D_1 = \Phi(s_2, s_1+2(n-1)+1)\Phi(s_1+2(n-1), s_1)\Phi(s_1-1, r)$$

[2]This will be achieved when agents exchange information about their estimates and "planned" weights simultaneously and set their actual weights as the minimum of the planned weights; see [14] where such a coordination scheme is described in detail.

for $i = 2, \ldots, t-1$

$$D_i = \Phi(s_{i+1}, s_i + 2(n-1) + 1)\Phi(s_i + 2(n-1), s_i)$$
$$\text{and} \quad D_t = \Phi(k, s_t + 2(n-1) + 1)\Phi(s_t + 2(n-1), s_t)$$

where Φ is replaced by an identity matrix wherever the first parameter of Φ is smaller than the second. Note that

$$\Phi(k, r) = D_t \cdots D_1.$$

For each i from 1 to t, D_i is a product of two or three matrices. Because we assume that the event $G(s_i)$ occurs for each i, the second matrix of each D_i, $\Phi(s_i + 2(n-1), s_i)$, has all elements greater than or equal to $\gamma^{2(n-1)}$ by Lemma 2, i.e.,

$$\min_{l,j \in \mathcal{N}} [\Phi(s_i + 2(n-1), s_i)]_l^j \geq \gamma^{2(n-1)}.$$

This minimum element property remains when we multiply a matrix by any doubly stochastic matrix. Since $\Phi(\cdot, \cdot)$ is assumed to always be doubly stochastic, it follows that

$$\min_{l,j \in \mathcal{N}} [D_i]_l^j \geq \gamma^{2(n-1)} \quad \text{for all} \quad i \in \{1, \ldots, t\}.$$

Hence, the product of matrices $D_t \cdots D_1$ satisfies all the conditions of Lemma 5, with $\delta = \gamma^{2(n-1)}$. Therefore

$$\max_{(i,j) \in \{1,\ldots,n\}^2} \left| [D_t \cdots D_1]_i^j - \frac{1}{n} \right| \leq 2\left(1 + \frac{1}{\gamma^{2(n-1)}}\right)(1 - \gamma^{2(n-1)})^t.$$

Since $\Phi(k, r) = D_t \cdots D_1$, the left-hand side of the equation above is equal to the definition of $b(k, r)$. Thus, we obtain (16). ■

We use the preceding result to show that the expected value of the metric $b(k, s)$ decays geometrically in $k - s$, which is formalized in the next lemma.

Lemma 7: (Geometric Decay) Let Connectivity and Doubly Stochastic Weights Assumptions hold [cf. Assumptions 2 and 3]. Then

$$E[b(k, s)] \leq C\beta^{k-s} \quad \text{for all} \quad k \geq s \tag{17}$$

where β and C are given by

$$C = \left(3 + \frac{2}{\gamma^{2(n-1)}}\right)\exp\left\{-\frac{\gamma^{4(n-1)}}{2}\right\} \tag{18}$$
$$\beta = \exp\left\{-\frac{\gamma^{4(n-1)}}{4(n-1)}\right\}. \tag{19}$$

Proof: To obtain this result, we first divide the interval $s, \ldots, k$ into a number of intervals of length $2(n-1)$. We then proceed to use the independence of the events during these separate intervals to get (17).

Let the number of desired intervals of $s, \ldots, k$ be given by

$$t = \left\lfloor \frac{k - s + 1}{2(n-1)} \right\rfloor. \tag{20}$$

Let Z_i for $i = 1, \ldots, t$ be a sequence of independent Bernoulli random variables with success probability $\gamma^{2(n-1)}$. For each i, let the random variable Z_i be correlated with the realization A^∞ in the following way: if $Z_i = 1$, then the event $G(s + (i-1)2(n-1))$ occurs. Note that the events $G(s + (i-1)2(n-1))$ for different i's are independent, and, therefore, this construction is valid.

We condition the random variable $b(k, s)$ on $\sum_{i=1}^t Z_i$ to bound it's expected value

$$\begin{aligned} &E[b(k, s)] \\ &= E\left[b(k, s) \middle| \sum_{i=1}^t Z_i \geq \frac{\gamma^{2(n-1)}t}{2}\right] P\left(\sum_{i=1}^t Z_i \geq \frac{\gamma^{2(n-1)}t}{2}\right) \\ &+ E\left[b(k, s) \middle| \sum_{i=1}^t Z_i < \frac{\gamma^{2(n-1)}t}{2}\right] P\left(\sum_{i=1}^t Z_i < \frac{\gamma^{2(n-1)}t}{2}\right). \end{aligned}$$

Since all the terms in the right-hand side of the equation above are smaller than 1, the following bound holds:

$$E[b(k, s)] \leq E\left[b(k, s) \middle| \sum_{i=1}^t Z_i \geq \frac{\gamma^{2(n-1)}t}{2}\right] + P\left(\sum_{i=1}^t Z_i < \frac{\gamma^{2(n-1)}t}{2}\right). \tag{21}$$

To complete this lemma, we separately bound the two terms on the right-hand side of the equation above. By Lemma 6, we get that if more than $\gamma^{2(n-1)}t/2$ events of the form $G(s + (i-1)2(n-1))$ occur then

$$\begin{aligned} b(k, s) &\leq 2\left(1 + \frac{1}{\gamma^{2(n-1)}}\right)(1 - \gamma^{2(n-1)})^{(\gamma^{2(n-1)}/2)t} \\ &= 2\left(1 + \frac{1}{\gamma^{2(n-1)}}\right)e^{\ln(1-\gamma^{2(n-1)})(\gamma^{2(n-1)}/2)t} \\ &\leq 2\left(1 + \frac{1}{\gamma^{2(n-1)}}\right)e^{-(\gamma^{4(n-1)}/2)t} \end{aligned}$$

where the last inequality follows from $\ln(1 - z) \leq -z$ for all $z < 1$. By integrating over all possible events that satisfy $\sum_{i=1}^t Z_i \geq (\gamma^{2(n-1)}t/2)$

$$E\left[b(k, s) \middle| \sum_{i=1}^t Z_i \geq \frac{\gamma^{2(n-1)}t}{2}\right] \leq \left(2 + \frac{2}{\gamma^{2(n-1)}}\right)e^{-(\gamma^{4(n-1)}/2)t}. \tag{22}$$

From large deviation theory, we can use Hoeffding's inequality to bound $P\left(\sum_{i=1}^t Z_i < (\gamma^{2(n-1)}t/2)\right)$. From Hoeffding's inequality, we get that for any u

$$P\left(\sum_{i=1}^t Z_i < \gamma^{2(n-1)}t - ut\right) \leq e^{-2u^2 t}.$$

By letting $u = (\gamma^{2(n-1)}/2)$ we obtain

$$P\left(\sum_{i=1}^t Z_i < \frac{\gamma^{2(n-1)}t}{2}\right) \leq e^{-(\gamma^{4(n-1)}/2)t}$$

which combined with (21) and (22), produces

$$E[b(k,s)] \le \left(3 + \frac{2}{\gamma^{2(n-1)}}\right) e^{-(\gamma^{4(n-1)}/2)t}. \qquad (23)$$

From (20), we can construct the following bound on t:

$$t \ge \frac{k-s+1}{2(n-1)} + 1 \ge \frac{k-s}{2(n-1)} + 1. \qquad (24)$$

By using the bound of (24) for the value of t in (23), we obtain

$$E[b(k,s)] \le \left(3 + \frac{2}{\gamma^{2(n-1)}}\right) \exp\left\{-\frac{\gamma^{4(n-1)}}{4(n-1)}(k-s) - \frac{\gamma^{4(n-1)}}{2}\right\}$$

which completes the lemma. ■

The preceding lemma establishes that for all $k \ge s, E[b(k,s)]$ decays exponentially in $k-s$. Combined with the results of Section V, this will enable us to analyze the iterates of the distributed subgradient method.

V. Analysis of the Subgradient Method

In this section, we study the convergence behavior of the iterates $x_i(k)$ of the distributed subgradient method given in (3). We start by analyzing the asymptotic disagreement in the iterates (or agent estimates). We provide uniform upper bounds on the "disagreement in agent estimates" that hold at each iteration and for any stepsize sequence. We also establish almost sure agreement in the limit under some assumptions on the stepsize sequence. We then analyze the convergence of agent estimates to the optimal solution of problem (2).

A. Disagreement in Agent Estimates

We first study the asymptotic disagreement in agent estimates. Using the linearity of the update rule given in (3) and the definition of the transition matrices [cf. (6)], we have shown that the iterates generated by this method satisfy the following relation: for any $i \in \mathcal{N}$, and any s and k with $k \ge s \ge 0$

$$x_i(k+1) = \sum_{j=1}^{n} [\Phi(k,s)]_{ij} x_j(s) - \sum_{r=s+1}^{k} \sum_{j=1}^{n} [\Phi(k,r)]_{ij} \alpha(r-1) d_j(r-1) - \alpha(k) d_i(k) \qquad (25)$$

[cf. (7)].

To analyze the disagreement in the iterates $\{x_i(k)\}$ for all $i \in \mathcal{N}$, we find it useful to introduce a related sequence $\{y(k)\}$, with $y(k) \in \mathbb{R}^m$ for all $k \ge 0$, defined as follows: Let the initial iterate $y(0)$ be given by

$$y(0) = \frac{1}{n} \sum_{j=1}^{n} x_j(0). \qquad (26)$$

At time $k+1$, the iterate $y(k+1)$ is obtained by

$$y(k+1) = y(k) - \frac{\alpha(k)}{n} \sum_{j=1}^{n} d_j(k). \qquad (27)$$

Equivalently, for all $k \ge 0$, $y(k)$ is given by

$$y(k) = \frac{1}{n} \sum_{j=1}^{n} x_j(0) - \frac{1}{n} \sum_{s=1}^{k} \alpha(s) \sum_{j=1}^{n} d_j(s-1). \qquad (28)$$

The iterate $y(k)$ represents a centralized combination of all the information that has become available in the system by time k. Since the vector $d_j(k)$ denotes a subgradient of the agent j objective function $f_j(x)$ at $x = x_j(k)$, iteration (27) can be viewed as an *approximate subgradient method*, in which a subgradient at $x = x_j(k)$ is used instead of a subgradient at $x = y(k)$. Our goal is to provide bounds on the norm of the difference between $y(k)$ and $x_i(k)$, and use these bounds and the behavior of the approximate subgradient method to analyze the convergence of the estimates $x_i(k)$.

We adopt the following standard assumption in our analysis.

Assumption 4: (Bounded Subgradients) Assume there exists a scalar L such that for any $x \in \mathbb{R}^m$, any $j \in \mathcal{N}$, all subgradients $s \in \partial f_j(x)$ satisfy $\|s\| \le L$.

This assumption is satisfied, for example, when each f_i is polyhedral (i.e., f_i is the pointwise maximum of a finite number of affine functions). We also assume in the remainder of the paper

$$\max_{1 \le j \le n} \|x_j(0)\| \le L \qquad (29)$$

where $x_j(0)$ denotes the initial vector (estimate) of agent j. This assumption is for notational convenience and can be relaxed at the expense of additional terms in the estimates which do not change the asymptotic results.

The following proposition provides a uniform bound on the norm of the difference between $y(k)$ and $x_i(k)$ that holds for all $i \in \mathcal{N}$ and all $k \ge 0$. We also consider the (weighted) averaged-vectors $\tilde{x}_i(k)$ and $\tilde{y}(k)$ defined for all $k \ge 0$ as

$$\tilde{x}_i(k) = \frac{1}{\sum_{s=0}^{k} \alpha(s)} \sum_{t=0}^{k} \alpha(t) x_i(t) \quad \text{and} \qquad (30)$$

$$\tilde{y}(k) = \frac{1}{\sum_{s=0}^{k} \alpha(s)} \sum_{t=0}^{k} \alpha(t) y(t) \qquad (31)$$

and provide a bound on the norm of the difference between $\tilde{y}(k)$ and $\tilde{x}_i(k)$.

Proposition 1: Let Bounded Subgradients assumption hold [cf. Assumption 4]. Let the sequence $\{y(k)\}$ be generated by iteration (27), and the sequences $\{x_i(k)\}$ for $i \in \mathcal{N}$ be generated by iteration (3).

(a) For all $i \in \mathcal{N}$ and $k \ge 1$, an upper bound on $\|y(k) - x_i(k)\|$ is given by

$$\|y(k) - x_i(k)\| \le nL \sum_{s=0}^{k-1} \alpha(s-1) b(k-1,s) + 2\alpha(k-1)L$$

where we define $\alpha(-1) = 1$ for convenience.

(b) For all $i \in \mathcal{N}$ and $k \geq 1$, an upper bound on $\|\tilde{y}(k) - \tilde{x}_i(k)\|$ is given by

$$\|\tilde{y}(k) - \tilde{x}_i(k)\| \leq \frac{1}{\sum_{r=0}^{k} \alpha(r)} \sum_{t=0}^{k} \alpha(t) \left[nL \sum_{s=0}^{t-1} \alpha(s-1) b(t-1,s) + 2\alpha(t-1)L \right]$$

where we let $\sum_{s=0}^{-1}(\cdot) = 0$ for convenience.

Proof:

(a) Substituting $s = 0$ in (25), we obtain

$$x_i(k+1) = \sum_{j=1}^{n} [\Phi(k,0)]_{ij} x_j(0) - \sum_{r=1}^{k} \left(\sum_{j=1}^{n} [\Phi(k,r)]_{ij} \alpha(r-1) d_j(r-1) \right) - \alpha(k) d_i(k).$$

Subtracting the preceding relation from (28) and taking the norm, we obtain for all $k \geq 1$ and $i \in \mathcal{N}$

$$\|y(k) - x_i(k)\| \leq \left\| \sum_{j=1}^{n} x_j(0) \left(\frac{1}{n} - [\Phi(k-1,0)]_{ij} \right) - \sum_{s=1}^{k-1} \alpha(s-1) \sum_{j=1}^{n} \left(\frac{1}{n} - [\Phi(k-1,s)]_j^i \right) d_j(s-1) - \alpha(k-1) \left(\frac{1}{n} \sum_{j=1}^{n} d_j(k-1) - d_i(k-1) \right) \right\|.$$

Therefore, for all $k \geq 1$ and $i \in \mathcal{N}$

$$\|y(k) - x_i(k)\| \leq \max_{1 \leq j \leq n} \|x_j(0)\| \sum_{j=1}^{n} \left| \frac{1}{n} - [\Phi(k-1,0)]_{ij} \right| + \sum_{s=1}^{k-1} \alpha(s-1) \sum_{j=1}^{n} \|d_j(s-1)\| \left| \frac{1}{n} - [\Phi(k-1,s)]_{ij} \right| + \frac{\alpha(k-1)}{n} \sum_{j=1}^{n} \|d_j(k-1) - d_i(k-1)\|.$$

Using the assumption that $\max_{1 \leq j \leq n} \|x_j(0)\| \leq L$, the Bounded Subgradients assumption [cf. Assumption 4], and the definition

$$b(k,s) = \max_{i,j} \left| \frac{1}{n} - [\Phi(k,s)]_{ij} \right|$$

[cf. (15)], it follows from the preceding relation that for all $i \in \mathcal{N}$ and $k \geq 1$

$$\|y(k) - x_i(k)\| \leq nL \sum_{s=0}^{k-1} \alpha(s-1) b(k-1,s) + 2\alpha(k-1)L$$

where we used $\alpha(-1) = 1$. This establishes part (a).

(b) Using the definition of the averaged-vectors in (30), (31), we obtain for all $i \in \mathcal{N}$ and $k \geq 1$

$$\begin{aligned} \|\tilde{y}(k) - \tilde{x}_i(k)\| &= \left\| \frac{1}{\sum_{s=0}^{k} \alpha(s)} \sum_{t=0}^{k} \alpha(t)(y(t) - x_i(t)) \right\| \\ &\leq \frac{1}{\sum_{s=0}^{k} \alpha(s)} \sum_{t=0}^{k} \alpha(t) \|y(t) - x_i(t)\| \\ &= \frac{1}{\sum_{s=0}^{k} \alpha(s)} \left[\|y(0) - x_i(0)\| + \sum_{t=1}^{k} \alpha(t) \|y(t) - x_i(t)\| \right]. \end{aligned} \tag{32}$$

Since $y(0)$ is the average of $x_j(0)$ for all $j \in \mathcal{N}$ and $\|x_j(0)\| \leq L$, [cf. (26) and (29)]

$$\|y(0) - x_i(0)\| \leq \frac{1}{n} \sum_{j=1}^{n} \|x_j(n)\| + \|x_i(0)\| \leq 2L = 2\alpha(-1)L.$$

Using this bound in (32)

$$\|\tilde{y}(k) - \tilde{x}_i(k)\| \leq \frac{1}{\sum_{s=0}^{k} \alpha(s)} \left[2\alpha(-1)L + \sum_{t=1}^{k} \alpha(t) \|y(t) - x_i(t)\| \right].$$

Using the estimate in part (a) for $t = 1, \ldots, k$ and the convention that $\sum_{s=0}^{-1}(\cdot) = 0$ for $t = 0$, we obtain

$$\|\tilde{y}(k) - \tilde{x}_i(k)\| \leq \frac{1}{\sum_{r=0}^{k} \alpha(r)} \sum_{t=0}^{k} \alpha(t) \left[nL \sum_{s=0}^{t-1} \alpha(s-1) b(t-1,s) + 2\alpha(t-1)L \right]$$

which completes the proof. ∎

We next study the almost sure convergence properties of the sequences $\{\|y(k) - x_i(k)\|\}$ under some additional assumptions on the stepsize sequence $\{\alpha(k)\}$. We rely on the following standard convergence result for sequences of random variables, which is an immediate consequence of the supermartingale convergence theorem (see Bertsekas and Tsitsiklis [2]).

Lemma 8: Consider a probability space (Ω, F, P) and let $\{F(k)\}$ be an increasing sequence of σ-fields contained in F. Let $\{V(k)\}$ and $\{Z(k)\}$ be sequences of nonnegative random variables (with finite expectation) adapted to $\{F(k)\}$ that satisfy

$$E[V(k+1) \mid F(k)] \leq V(k) + Z(k)$$
$$\sum_{k=1}^{\infty} E[Z(k)] < \infty.$$

Then, $V(k)$ converges with probability one, as $k \to \infty$.

The following lemma on the infinite sum of products of the components of two sequences will also be used in establishing our convergence results (see Lemma 7 in [15] for the proof).

Lemma 9: Let $0 < \beta < 1$ and let $\{\gamma(k)\}$ be a positive scalar sequence. Assume that $\lim_{k\to\infty} \gamma(k) = 0$. Then

$$\lim_{k\to\infty} \sum_{\ell=0}^{k} \beta^{k-\ell}\gamma(\ell) = 0.$$

In addition, if $\sum_{k=1}^{\infty} \gamma(k) < \infty$, then

$$\sum_{k=1}^{\infty} \sum_{\ell=0}^{k} \beta^{k-\ell}\gamma(\ell) < \infty.$$

The next proposition shows that under some assumptions on the stepsize, the sequences $\{\|y(k) - x_i(k)\|\}$ converge to zero with probability one, thus establishing almost sure agreement among agent estimates in the limit.

Proposition 2: Let Connectivity, Doubly Stochastic Weights, and Bounded Subgradients assumptions hold [cf. Assumptions 2, 3, and 4]. Let the sequence $\{y(k)\}$ be generated by iteration (27), and the sequences $\{x_i(k)\}$ for $i \in \mathcal{N}$ be generated by iteration (3). Assume that the stepsize sequence satisfies $\sum_{k=1}^{\infty} \alpha(k)^2 < \infty$. Then, for all $i \in \mathcal{N}$, we have

(a) $\sum_{k=1}^{\infty} \alpha(k)\|y(k) - x_i(k)\| < \infty$ with probability 1.

(b) $\lim_{k\to\infty} \|y(k) - x_i(k)\| = 0$ with probability 1.

Proof:

(a) By multiplying the relation in Proposition 1(a) with $\alpha(k)$, we obtain

$$\alpha(k)\|y(k) - x_i(k)\| \leq nL\sum_{s=0}^{k-1} \alpha(k)\alpha(s-1)b(k-1,s) + 2\alpha(k)\alpha(k-1)L.$$

Taking the expectation and using the estimate from Lemma 7, i.e.,

$$E[b(k,s)] \leq C\beta^{k-s} \quad \text{for all } k \geq s$$

where $0 < \beta < 1$ and $C \geq 0$ are given by (18), (19), we have

$$E[\alpha(k)\|y(k) - x_i(k)\|] \leq nLC\sum_{s=0}^{k-1} \alpha(k)\alpha(s-1)\beta^{k-1-s} + 2\alpha(k)\alpha(k-1)L.$$

Using the relations $\alpha(k)\alpha(s-1) \leq \alpha(k)^2 + \alpha(s-1)^2$ and $2\alpha(k)\alpha(k-1) \leq \alpha(k)^2 + \alpha(k-1)^2$ for any k and s, this implies that

$$E[\alpha(k)\|y(k) - x_i(k)\|] \leq \frac{nLC}{1-\beta}\alpha(k)^2 + nLC\sum_{s=0}^{k-1} \alpha(s-1)^2\beta^{k-1-s} + L(\alpha(k)^2 + \alpha(k-1)^2).$$

Summing over $k \geq 1$ and grouping some of the terms, we obtain

$$\sum_{k=1}^{\infty} E[\alpha(k)\|y(k) - x_i(k)\|] \leq nLC\sum_{k=1}^{\infty}\sum_{s=0}^{k-1} \alpha(s-1)^2\beta^{k-1-s} + \sum_{k=1}^{\infty}\left(\left(\frac{nLC}{1-\beta} + L\right)\alpha(k)^2 + L\alpha(k-1)^2\right).$$

In this relation, the first term is summable since $\sum_k \alpha(k)^2 < \infty$ and the second term is summable by Lemma 9, showing that

$$\sum_{k=1}^{\infty} E[\alpha(k)\|y(k) - x_i(k)\|] < \infty.$$

By the monotone convergence theorem, this implies that

$$E\left[\sum_{k=1}^{\infty} \alpha(k)\|y(k) - x_i(k)\|\right] < \infty$$

and therefore

$$\sum_{k=1}^{\infty} \alpha(k)\|y(k) - x_i(k)\| < \infty \qquad \text{with probability } 1.$$

(b) Using the iterations (3) and (27), we obtain for all $k \geq 1$ and $i \in \mathcal{N}$

$$y(k+1) - x_i(k+1) = \left(y(k) - \sum_{j=1}^{n} a_{ij}(k)x_j(k)\right) - \alpha(k)\left(\frac{1}{n}\sum_{j=1}^{n} d_j(k) - d_i(k)\right).$$

Therefore, using the stochasticity of the weights $a_{ij}(k)$ and the subgradient boundedness, we obtain

$$\|y(k+1) - x_i(k+1)\| \leq \sum_{j=1}^{n} a_{ij}(k)\|y(k) - x_j(k)\| + 2L\alpha(k).$$

Taking the square of both sides and using the convexity of the squared-norm function $\|\cdot\|^2$, this yields

$$\|y(k+1) - x_i(k+1)\|^2 \leq \sum_{j=1}^{n} a_{ij}(k)\|y(k) - x_j(k)\|^2 + 4L\alpha(k)\sum_{j=1}^{n} a_{ij}(k)\|y(k) - x_j(k)\| + 4L^2\alpha(k)^2.$$

Summing over all i and using the doubly stochasticity of the weights $a_{ij}(k)$ (i.e., $\sum_{j=1}^{n} a_{ij}(k) = 1$ for all i), we have for all $k \geq 1$

$$\sum_{i=1}^{n} \|y(k+1) - x_i(k+1)\|^2 \leq sum_{i=1}^{n}\|y(k) - x_i(k)\|^2 + 4L\alpha(k)\sum_{i=1}^{n} \|y(k) - x_i(k)\| + 4L^2n\alpha(k)^2.$$

By part (a) of this lemma, we have $\sum_{k=1}^{\infty}\alpha(k)\|y(k)-x_i(k)\| < \infty$ with probability 1. Since, we also have $\sum_k \alpha(k)^2 < \infty$, Lemma 8 applies and implies that $\sum_{i=1}^n \|y(k)-x_i(k)\|^2$ converges with probability 1, as $k \to \infty$.

We next show that the sequence $\|y(k)-x_i(k)\|$ converges to zero with probability 1 for all $i \in \mathcal{N}$. Taking the expectation in the relation in Proposition 1(a) and using the estimate from Lemma 7, we obtain

$$E[\|y(k)-x_i(k)\|] \le nLC\sum_{s=0}^{k-1}\alpha(s-1)\beta^{k-1-s}+2\alpha(k-1)L.$$

Since $\alpha(k) \to 0$ as $k \to \infty$, Lemma 9 implies that $\lim_{k\to\infty}\sum_{s=0}^{k-1}\alpha(s-1)\beta^{k-1-s} = 0$. Therefore, taking the limit inferior in the preceding relation and using Fatou's Lemma (which applies since the random variables $\|y(k)-x_i(k)\|$ are nonnegative for all i and k), we obtain

$$\begin{aligned}0 &\le E\left[\liminf_{k\to\infty}\|y(k)-x_i(k)\|\right]\\ &\le \liminf_{k\to\infty} E[\|y(k)-x_i(k)\|] \le 0.\end{aligned}$$

Thus, the nonnegative random variable $\liminf_{k\to\infty}\|y(k)-x_i(k)\|$ has expectation 0, which implies that

$$\liminf_{k\to\infty}\|y(k)-x_i(k)\| = 0 \qquad \text{with probability } 1.$$

Since $\sum_{i=1}^n \|y(k)-x_i(k)\|^2$ converges with probability 1, as $k\to\infty$, this implies that for all $i \in \mathcal{N}$

$$\lim_{k\to\infty}\|y(k)-x_i(k)\| = 0 \qquad \text{with probability } 1$$

completing the proof. ■

B. Convergence of Agent Estimates

This section studies the convergence of the agent estimates to the optimal solution of problem (2). We first establish a relation for the squared-distance of the iterates $y(k)$ to the optimal solution set X^*, which will be key in the convergence analysis of the distributed subgradient method. This relation was proven in [14], and, therefore, the proof is omitted. In the following lemma and thereafter, we use the notation $f(x)=\sum_{i=1}^n f_i(x)$.

Lemma 10: Let the sequence $\{y(k)\}$ be generated by iteration (27), and the sequences $\{x_i(k)\}$ for $i \in \mathcal{N}$ be generated by iteration (3). Let $\{g_i(k)\}$ be a sequence of subgradients such that $g_i(k) \in \partial f_i(y(k))$ for all $i \in \mathcal{N}$ and $k \ge 0$. We then have for all $k \ge 0$ and any $x^* \in X^*$

$$\begin{aligned}\|y(k+1)-x^*\|^2 &\le \|y(k)-x^*\|^2\\ &+\frac{2\alpha(k)}{n}\sum_{j=1}^n(\|d_j(k)\|+\|g_j(k)\|)\|y(k)-x_j(k)\|\\ &-\frac{2\alpha(k)}{n}[f(y(k))-f^*]+\frac{\alpha^2(k)}{n^2}\sum_{j=1}^n\|d_j(k)\|^2.\end{aligned}$$

The next proposition establishes upper bounds on the difference of the objective function value of the averaged iterates [$\tilde{y}(k)$ and $\tilde{x}(k)$] from the optimal value f^*. It relies on combining the bounds on the difference between the iterates given in Proposition 1 with the preceding lemma.

Proposition 3: Let Bounded Subgradients assumption hold [cf. Assumption 4]. Let the sequence $\{y(k)\}$ be generated by iteration (27), and the sequences $\{x_i(k)\}$ for $i \in \mathcal{N}$ be generated by iteration (3).

(a) Let $\tilde{y}(k)$ be the averaged vector defined in (31). An upper bound on the objective function $f(\tilde{y}(k))$ is given by

$$\begin{aligned}f(\tilde{y}(k)) \le f^* + \frac{n}{2\sum_{r=0}^k\alpha(r)}\Bigg(&L^2\sum_{t=0}^k\alpha^2(t)+\\ &4L\sum_{t=0}^k\alpha(t)\left[nL\sum_{s=0}^{t-1}\alpha(s-1)b(t-1,s)\right.\\ &\left.+2\alpha(t-1)L\right]+dist^2(y(0),X^*)\Bigg).\end{aligned}$$

(b) Let $\tilde{x}_i(k)$ be the averaged vector defined in (30). An upper bound on the objective value $f(\tilde{x}_i(k))$ for each i is given by

$$\begin{aligned}f(\tilde{x}_i(k)) \le f^* + \frac{n}{2\sum_{r=0}^k\alpha(r)}\Bigg(&L^2\sum_{t=0}^k\alpha^2(t)\\ &+6L\sum_{t=0}^k\alpha(t)\left[nL\sum_{s=0}^{t-1}\alpha(s-1)b(t-1,s)\right.\\ &\left.+2\alpha(t-1)L\right]+dist^2(y(0),X^*)\Bigg).\end{aligned}$$

Proof:

(a) By using Lemma 10 and the Bounded Subgradients assumption [cf. Assumption 4], we have for all $t \ge 1$

$$\begin{aligned}dist^2(y(t+1),X^*) &\le dist^2(y(t),X^*)+\frac{\alpha^2(t)L^2}{n}\\ &+\frac{4L\alpha(t)}{n}\sum_{j=1}^n\|y(t)-x_j(t)\|-\frac{2\alpha(t)}{n}[f(y(t))-f^*].\end{aligned}$$

Summing the preceding relation for $t = 0,\ldots,k$, we obtain for $k \ge 0$

$$\begin{aligned}dist^2(y(k+1),X^*) &\le dist^2(y(0),X^*)+\frac{L^2}{n}\sum_{t=0}^k\alpha^2(t)\\ &+\frac{4L}{n}\sum_{t=0}^k\alpha(t)\sum_{j=1}^n\|y(t)-x_j(t)\|-\frac{2}{n}\sum_{t=0}^k\alpha(t)[f(y(t))-f^*].\end{aligned}$$

Since $dist^2(y(k+1),X^*) \ge 0$, this yields

$$\begin{aligned}0 \le dist^2(y(0),X^*)&+\frac{4L}{n}\sum_{t=0}^k\alpha(t)\sum_{j=1}^n\|y(t)-x_j(t)\|\\ &-\frac{2}{n}\sum_{t=0}^k\alpha(t)[f(y(t))-f^*]+\frac{L^2}{n}\sum_{t=0}^k\alpha^2(t).\end{aligned}$$

Using the estimate from part (a) of Proposition 1, we obtain

$$0 \le dist^2(y(0), X^*) + \frac{L^2}{n}\sum_{t=0}^{k}\alpha^2(t)$$
$$+ 4L\sum_{t=0}^{k}\alpha(t)\left[nL\sum_{s=0}^{t-1}\alpha(s-1)b(t-1,s) + 2\alpha(t-1)L\right]$$
$$- \frac{2}{n}\sum_{t=0}^{k}\alpha(t)[f(y(t)) - f^*].$$

Multiplying this relation by $n/2\sum_{r=0}^{k}\alpha(r)$, we obtain

$$0 \le \frac{n}{2\sum_{r=0}^{k}\alpha(r)}\left(dist^2(y(0), X^*) + L^2\sum_{t=0}^{k}\alpha^2(t)\right.$$
$$\left.+ 4L\sum_{t=0}^{k}\alpha(t)\left[nL\sum_{s=0}^{t-1}\alpha(s-1)b(t-1,s) + 2\alpha(t-1)L\right]\right)$$
$$- \frac{1}{\sum_{r=0}^{k}\alpha(r)}\sum_{t=0}^{k}\alpha(t)f(y(t)) + f^*. \tag{33}$$

By the convexity of the function f, we have

$$f(\tilde{y}(k)) \le \frac{1}{\sum_{r=0}^{k}\alpha(r)}\sum_{t=0}^{k}\alpha(t)f(y(t))$$

where $\tilde{y}(k) = (1/\sum_{r=0}^{k}\alpha(r))\sum_{t=0}^{k}\alpha(t)y(t)$. Combining this relation with (33) yields the desired result.

(b) We next prove the estimate for $f(\tilde{x}_i(k))$. Using the subgradient definition for the averaged-vectors $\tilde{x}_i(k)$, we have for all $i \in \mathcal{N}$ and all $k \ge 0$

$$f(\tilde{x}_i(k)) \le f(\tilde{y}(k)) + \sum_{j=1}^{n}\tilde{g}_{ij}(k)'(\tilde{x}_i(k) - \tilde{y}(k))$$

where $\tilde{g}_{ij}(k)$ is a subgradient of the objective function f_j at $\tilde{x}_i(k)$. Since by assumption $\|\tilde{g}_{ij}(k)\| \le L$ for all i, $j \in \mathcal{N}$, and $k \ge 0$, it follows that

$$f(\tilde{x}_i(k)) \le f(\tilde{y}(k)) + nL\|\tilde{x}_i(k) - \tilde{y}(k)\|.$$

Using the estimate in part (a) and part (b) of Proposition 1, we obtain for all $i \in \mathcal{N}$ and $k \ge 0$

$$f(\tilde{x}_i(k)) \le f^* + \frac{n}{2\sum_{r=0}^{k}\alpha(r)}\left(dist^2(y(0), X^*)\right.$$
$$+ 6L\sum_{t=0}^{k}\alpha(t)\left[nL\sum_{s=0}^{t-1}\alpha(s-1)b(t-1,s) + 2\alpha(t-1)L\right]$$
$$\left.+ L^2\sum_{t=0}^{k}\alpha^2(t)\right),$$

which yields the desired result. ■

We use the previous two lemmas to study the convergence of the iterates of the distributed subgradient method under two stepsize rules: a diminishing stepsize rule, whereby the stepsize sequence $\{\alpha(k)\}$ satisfies $\sum_{k=0}^{\infty}\alpha(k) = \infty$ and $\sum_{k=0}^{\infty}\alpha(k)^2 < \infty$, and a constant stepsize rule, whereby the the stepsize sequence $\{\alpha(k)\}$ is such that $\alpha(k) = \alpha$ for some constant $\alpha > 0$ and all k.

The next theorem contains our main convergence result for the diminishing stepsize rule.

Theorem 1: Let Connectivity, Doubly Stochastic Weights, and Bounded Subgradients assumptions hold [cf. Assumptions 2, 3, and 4]. Let the sequences $\{x_i(k)\}$ for $i \in \mathcal{N}$ be generated by iteration (3) with the stepsize satisfying $\sum_{k=0}^{\infty}\alpha(k) = \infty$ and $\sum_{k=0}^{\infty}\alpha(k)^2 < \infty$. Then, there exists an optimal point $z^* \in X^*$ such that for all $i \in \mathcal{N}$

$$\lim_{k\to\infty} x_i(k) = z^* \quad \text{with probability } 1.$$

Proof: From Lemma 10 and using the subgradient boundedness, we have for all $k \ge 0$ and any $x^* \in X^*$

$$\|y(k+1) - x^*\|^2 + \frac{2\alpha(k)}{n}[f(y(k)) - f^*] \le \|y(k) - x^*\|^2 +$$
$$\frac{4L\alpha(k)}{n}\sum_{j=1}^{n}\|y(k) - x_j(k)\| + \frac{L^2\alpha^2(k)}{n}. \tag{34}$$

Summing the preceding relation over $k \ge 1$, we obtain

$$\sum_{k=1}^{\infty}\frac{2\alpha(k)}{n}[f(y(k)) - f^*] \le \|y(1) - x^*\|^2 +$$
$$\frac{4L}{n}\sum_{k=1}^{\infty}\alpha(k)\sum_{j=1}^{n}\|y(k) - x_j(k)\| + \frac{L^2}{n}\sum_{k=1}^{\infty}\alpha^2(k).$$

Using $\sum_{k=1}^{\infty}\alpha(k)\sum_{j=1}^{n}\|y(k) - x_j(k)\| < \infty$ with probability 1 (cf. Proposition 2) and the assumption $\sum_k \alpha(k)^2 < \infty$, it follows that

$$\sum_{k=1}^{\infty}\alpha(k)[f(y(k)) - f^*] < \infty \quad \text{with probability } 1.$$

Together with $f(y(k)) \ge f^*$ and the assumption $\sum_k \alpha(k) = \infty$, this implies that

$$\liminf_{k\to\infty} f(y(k)) = f^* \quad \text{with probability } 1. \tag{35}$$

We next show that each sequence $\{x_i(k)\}$ converges to the same optimal point. By dropping the nonnegative term $2\alpha(k)/n[f(y(k)) - f^*]$ in (34), we obtain

$$\|y(k+1) - x^*\|^2 \le$$
$$\|y(k) - x^*\|^2 + \frac{4L\alpha(k)}{n}\sum_{j=1}^{n}\|y(k) - x_j(k)\| + \frac{L^2\alpha^2(k)}{n}.$$

We have $\sum_{k=1}^{\infty}\alpha(k)\|y(k) - x_i(k)\| < \infty$ with probability 1 from Proposition 2(a) and $\sum_k \alpha(k)^2 < \infty$ by assumption. Therefore, it follows from Lemma 8 that the sequence $\|y(k) - x^*\|^2$ converges with probability 1 for every $x^* \in X^*$. Since $y(k)$ is bounded, it must have a limit point, and in view of (35) and the continuity of f (due to convexity of f over $\mathbb{R}^n$), one of the limit points of $\{y(k)\}$ must belong to X^*; denote this limit point by z^*. Since the sequence $\|y(k) - z^*\|^2$ is convergent, it follows that $y(k)$ can have a unique limit point, i.e.,

$\lim_{k\to\infty} y(k) = z^*$ with probability 1. Together with Proposition 2(b), i.e., $\lim_{k\to\infty} \|y(k) - x_i(k)\| = 0$ with probability 1 for all $i \in \mathcal{N}$, this implies that every sequence $x_i(k)$ converges to the same $z^* \in X^*$ with probability 1. ■

Our final result concerns the convergence properties of the averaged iterates $\tilde{x}_i(k)$ under a constant stepsize rule.

Theorem 2: Let Connectivity, Doubly Stochastic Weights and Bounded Subgradients assumptions hold [cf. Assumptions 2, 3 and 4]. Assume also that for some constant α, $\alpha(k) = \alpha$ for all $k \geq 0$. Then, for all $j \in \mathcal{N}$ and all $k \geq 0$

$$\limsup_{k\to\infty} |f(\tilde{x}_i(k)) - f^*| \leq nL^2 \left(\frac{13}{2} + \frac{3n\alpha C}{1-\beta}\right) \tag{36}$$

with probability 1.

Proof: Proposition 3(b) provides the following bound for all $i \in \mathcal{N}$ and all $k \geq 0$

$$f(\tilde{x}_i(k)) \leq f^* + \frac{3n^2L^2\alpha}{(k+1)} \sum_{t=0}^{k}\sum_{s=0}^{t-1} b(t-1,s) + \frac{13}{2}nL^2\alpha + \frac{n\, dist^2(y(0),X^*)}{2(k+1)\alpha} \tag{37}$$

once we replace $\alpha(k)$ with a constant α. We can simplify the double sum in the equation above since $\sum_{s=0}^{-1}(\cdot)$ is defined to be equal to 0 [cf. Prop. 1(b)]

$$\frac{1}{(k+1)}\sum_{t=0}^{k}\sum_{s=0}^{t-1} b(t-1,s) = \frac{1}{(k+1)}\sum_{t=0}^{k-1}\sum_{s=0}^{t} b(t,s).$$

Taking the limit superior of both sides of (37) as k goes to infinity, we obtain

$$\limsup_{k\to\infty} f(\tilde{x}_i(k)) \leq f^* + \frac{13}{2}nL^2 + 3n^2L^2\alpha \limsup_{k\to\infty}\frac{1}{(k+1)}\sum_{t=0}^{k-1}\sum_{s=0}^{t} b(t,s).$$

Since $f(x) - f^* \geq 0$ for any $x \in \mathbb{R}^m$

$$\limsup_{k\to\infty} |f(\tilde{x}_i(k)) - f^*| \leq \frac{13}{2}nL^2 + 3n^2L^2\alpha \limsup_{k\to\infty}\frac{1}{(k+1)}\sum_{t=0}^{k-1}\sum_{s=0}^{t} b(t,s).$$

To complete this proof, we need to show that

$$\limsup_{k\to\infty}\frac{1}{(k+1)}\sum_{t=0}^{k-1}\sum_{s=0}^{t} b(t,s) \leq \frac{C}{1-\beta} \quad \text{with probability } 1. \tag{38}$$

To construct the bound of (38), we decompose the double sum above into averages of independent, identically distributed random variables. Note that $b(t,s)$ is identically distributed to $b(t',s')$ if $t-s = t'-s'$. Therefore, we rewrite

$$\frac{1}{(k+1)}\sum_{t=0}^{k-1}\sum_{s=0}^{t} b(t,s) = \sum_{r=0}^{k-1}\frac{1}{(k+1)}\sum_{t=r}^{k-1} b(t,t-r) \tag{39}$$

where all the terms with the same index r have the same distribution. From the construction of $b(\cdot,\cdot)$, we have that if $t \geq s > t' \geq s'$, then $b(t,s)$ is independent of $b(t',s')$. To exploit this independence, we rephrase the terms of (39) for any given $k \geq 1$ and $r \in \{0,\ldots,k-1\}$

$$\frac{1}{(k+1)}\sum_{t=r}^{k-1} b(t,t-r) = \frac{1}{(k+1)}\sum_{w=0}^{r}\sum_{t=r}^{k-1} b(t,t-r) I\{t \bmod (r+1) = w\} \tag{40}$$

where the mod operator determines the remainder of a division. In the double sum of (40), all the terms with the same r and w are independent and identically distributed. In particular, for any given k, r and w, the sum $\sum_{t=r}^{k-1} b(t,t-r)I\{t \bmod (r+1) = w\}$ includes at most $\lceil k-r/r+1\rceil$ terms, all independent and distributed according to $b(r,0)$. Therefore, we can construct a family of random variables $Y_{k,r,w}$ such that

$$\frac{1}{(k+1)}\sum_{t=r}^{k-1} b(t,t-r) \leq \frac{1}{(k+1)}\sum_{w=0}^{r} Y_{k,r,w} \tag{41}$$

and $Y_{k,r,w}$ is the sum of exactly $\lceil k-r/r+1\rceil$ independent terms distributed according to $b(r,0)$. The variable $b(r,0)$ takes value in $[0,1]$ and satisfies $E[b(r,0)] \leq C\beta^r$ from Lemma 7. Using Hoeffding's inequality, we obtain that for any positive constant $z_{k,r}$

$$\begin{aligned} P\left(Y_{k,r,w} \geq \left\lceil\frac{k-r}{r+1}\right\rceil (C\beta^r + z_{k,r})\right) &\leq P\left(Y_{k,r,w} \geq \left\lceil\frac{k-r}{r+1}\right\rceil (E[Y_{k,r}] + z_{k,r})\right) \\ &\leq e^{-2\lceil k-r/r+1\rceil z_{k,r}^2}. \end{aligned}$$

Using a union bound, we obtain that for any k, r and any $z_{k,r} > 0$

$$\begin{aligned} P\left(\sum_{w=0}^{r} Y_{k,r,w} \geq (r+1)\left\lceil\frac{k-r}{r+1}\right\rceil (C\beta^r + z_{k,r})\right) &\leq \sum_{w=0}^{r} P\left(Y_{k,r,w} \geq \left\lceil\frac{k-r}{r+1}\right\rceil (C\beta^r + z_{k,r})\right) \\ &\leq (r+1)e^{-2\lceil k-r/r+1\rceil z_{k,r}^2}. \end{aligned}$$

Scaling the sum above by $1/k+1$, we obtain

$$P\left(\frac{1}{k+1}\sum_{w=0}^{r} Y_{k,r,w} \geq \left(\frac{r+1}{k+1}\right)\left\lceil\frac{k-r}{r+1}\right\rceil (C\beta^r + z_{k,r})\right) \leq (r+1)e^{-2\lceil k-r/r+1\rceil z_{k,r}^2}. \tag{42}$$

Note that

$$\left(\frac{r+1}{k+1}\right)\left\lceil\frac{k-r}{r+1}\right\rceil \leq \left(\frac{r+1}{k+1}\right)\left(\frac{k-r}{r+1}+1\right) = 1$$

and, therefore, (42) can be simplified to

$$P\left(\frac{1}{k+1}\sum_{w=0}^{r} Y_{k,r,w} \geq C\beta^r + z_{k,r}\right) \leq (r+1)e^{-2\lceil k-r/r+1\rceil z_{k,r}^2} \leq (r+1)e^{-2(k+1/r+1-1)z_{k,r}^2}.$$

We now select the family of terms $z_{k,r}$ in order to balance both sides of the equation above. By restricting ourselves to $z_{k,r} \leq 1$ for all k and r, we obtain

$$P\left(\frac{1}{k+1}\sum_{w=0}^{r} Y_{k,r,w} \geq C\beta^r + z_{k,r}\right) \leq e^2(r+1)e^{-2(k+1/r+1)z_{k,r}^2}.$$

Let $z_{k,r} = (k+1)^{-1/4}(r+1)^{-2}$. Then

$$P\left(\frac{1}{k+1}\sum_{w=0}^{r} Y_{k,r,w} \geq C\beta^r + \frac{1}{(k+1)^{1/4}(r+1)^2}\right) \leq e^2(r+1)e^{-2(k+1/r+1)(k+1)^{-1/2}(r+1)^{-4}} = e^2(r+1)e^{-2(k+1)^{1/2}/(r+1)^5}.$$

Consider the case where $(k+1) \geq (r+1)^{21}$. In this case, $(k+1)^{1/4} \geq (r+1)^{5+1/4}$ and, therefore

$$P\left(\frac{1}{k+1}\sum_{w=0}^{r} Y_{k,r,w} \geq C\beta^r + \frac{1}{(k+1)^{1/4}(r+1)^2}\right) \leq e^2(r+1)e^{-2(k+1)^{1/4}(r+1)^{1/4}}.$$

Using a union bound for all r from 0 to $\lfloor (k+1)^{1/21} - 1\rfloor$, we obtain

$$P\left(\sum_{r=0}^{\lfloor (k+1)^{1/21}-1\rfloor} \frac{1}{k+1}\sum_{w=0}^{r} Y_{k,r,w} \geq \sum_{r=0}^{\lfloor (k+1)^{1/21}-1\rfloor} C\beta^r + \frac{1}{(k+1)^{1/4}(r+1)^2}\right) \leq \sum_{r=0}^{\lfloor (k+1)^{1/21}-1\rfloor} e^2(r+1)e^{-2(k+1)^{1/4}(r+1)^{1/4}}$$

which can be relaxed to

$$P\left(\sum_{r=0}^{\lfloor (k+1)^{1/21}-1\rfloor} \frac{1}{k+1}\sum_{w=0}^{r} Y_{k,r,w} \geq \sum_{r=0}^{\infty} C\beta^r + \frac{1}{(k+1)^{1/4}(r+1)^2}\right) \leq \sum_{r=0}^{\infty} e^2(r+1)e^{-2(k+1)^{1/4}(r+1)^{1/4}}$$

yielding

$$P\left(\sum_{r=0}^{\lfloor (k+1)^{1/21}-1\rfloor} \frac{1}{k+1}\sum_{w=0}^{r} Y_{k,r,w} \geq \frac{C}{1-\beta} + \frac{\pi^2}{6}(k+1)^{-1/4}\right) \leq \sum_{r=0}^{\infty} e^2(r+1)e^{-2(k+1)^{1/4}(r+1)^{1/4}}.$$

This combined with (41) produces

$$P\left(\sum_{r=0}^{\lfloor (k+1)^{1/21}-1\rfloor} \frac{1}{(k+1)}\sum_{t=r}^{k-1} b(t,t-r) \geq \frac{C}{1-\beta} + \frac{\pi^2}{6}(k+1)^{-1/4}\right) \sum_{r=0}^{\infty} e^2(r+1)e^{-2(k+1)^{1/4}(r+1)^{1/4}}. \quad (43)$$

We next consider the case $(k+1) < (r+1)^{21}$. By the monotone convergence theorem, we have that for any k

$$E\left[\sum_{r=\lfloor (k+1)^{1/21}\rfloor}^{\infty} \frac{1}{(k+1)}\sum_{t=r}^{k-1} b(t,t-r)\right] = \sum_{r=\lfloor (k+1)^{1/21}\rfloor}^{\infty} \frac{1}{(k+1)}\sum_{t=r}^{k-1} E[b(t,t-r)] \leq \sum_{r=\lfloor (k+1)^{1/21}\rfloor}^{\infty} \frac{1}{(k+1)}\sum_{t=r}^{k-1} C\beta^r$$

where the inequality follows from Lemma 7. We can relax the inequality above to obtain

$$E\left[\sum_{r=\lfloor (k+1)^{1/21}\rfloor}^{\infty} \frac{1}{(k+1)}\sum_{t=r}^{k-1} b(t,t-r)\right] \leq \sum_{r=\lfloor (k+1)^{1/21}\rfloor}^{\infty} \frac{1}{(k+1)}\sum_{t=0}^{k} C\beta^r = \sum_{r=\lfloor (k+1)^{1/21}\rfloor}^{\infty} C\beta^r$$

which yields

$$E\left[\sum_{r=\lfloor (k+1)^{1/21}\rfloor}^{\infty} \frac{1}{(k+1)}\sum_{t=r}^{k-1} b(t,t-r)\right] \leq \frac{C}{1-\beta}\beta^{\lfloor (k+1)^{1/21}\rfloor} \leq \frac{C}{\beta(1-\beta)}\beta^{(k+1)^{1/21}}.$$

From Markov's inequality, we know that for any non-negative random variable X, $P(X \geq \sqrt{E[X]}) \leq \sqrt{E[X]}$. Applying it on the equation above, we obtain

$$P\left(\sum_{r=\lfloor (k+1)^{1/21}\rfloor}^{\infty} \frac{1}{(k+1)}\sum_{t=r}^{k-1} b(t,t-r) \geq \sqrt{\frac{C}{\beta(1-\beta)}}\beta^{1/2(k+1)^{1/21}}\right) d \leq \sqrt{\frac{C}{\beta(1-\beta)}}\beta^{1/2(k+1)^{1/21}}. \quad (44)$$

Using a union bound, we combine (43) and (44) to get for any k

$$P\left(\sum_{r=0}^{\infty} \frac{1}{(k+1)}\sum_{t=r}^{k-1} b(t,t-r) \geq \frac{C}{1-\beta} + \frac{\pi^2}{6}(k+1)^{-1/4} + \sqrt{\frac{C}{\beta(1-\beta)}}\beta^{1/2(k+1)^{1/21}}\right) \leq \sqrt{\frac{C}{\beta(1-\beta)}}\beta^{1/2(k+1)^{1/21}} + \sum_{r=0}^{\infty}\left(e^2(r+1)e^{-2(k+1)^{1/4}(r+1)^{1/4}}\right). \quad (45)$$

Using the Borel-Cantelli Lemma on (45), we get that

$$\sum_{r=0}^{\infty} \frac{1}{(k+1)} \sum_{t=r}^{k-1} b(t, t-r) \geq \frac{C}{1-\beta} + \frac{\pi^2}{6}(k+1)^{-1/4} + \sqrt{\frac{C}{\beta(1-\beta)}} \beta^{1/2(k+1)^{1/21}} \quad (46)$$

occurs only for finitely many k's if

$$\sum_{k=0}^{\infty} \sum_{r=0}^{\infty} \left(e^2 (r+1) e^{-2(k+1)^{1/4}(r+1)^{1/4}} \right) + \sqrt{\frac{C}{\beta(1-\beta)}} \beta^{1/2(k+1)^{1/21}} < \infty.$$

The two terms in the summation above can be upper bounded by integrals. The first term is finite since

$$\sum_{k=0}^{\infty} \sum_{r=0}^{\infty} e^2 (r+1) e^{-2(k+1)^{1/4}(r+1)^{1/4}} \leq \int_{k=0}^{\infty} \int_{r=0}^{\infty} e^2 (r+2) e^{-2(k+1)^{1/4}(r+1)^{1/4}} \, dr \, dk < 599$$

and the second one is finite as well for any given $0 < \beta < 1$ since

$$\sum_{k=1}^{\infty} \sqrt{\beta}^{k^{1/21}} \leq \int_{k=0}^{\infty} \sqrt{\beta}^{k^{1/21}} < \frac{10^{20}}{\ln^{21} \sqrt{\beta}}.$$

Therefore, we obtain that with probability 1, (46) holds only for finitely many k's. Thus

$$\limsup_{k \to \infty} \frac{1}{k+1} \sum_{t=0}^{k-1} \sum_{s=0}^{t} b(t, s) \leq \lim_{k \to \infty} \frac{C}{1-\beta} + \frac{\pi^2}{6}(k+1)^{-1/4} + \sqrt{\frac{C}{\beta(1-\beta)}} \beta^{1/2(k+1)^{1/21}} = \frac{C}{1-\beta}$$

proving (38). ■

Our last result thus shows that with a constant stepsize rule, we can bound (with probability 1) the difference of the objective function value of the iterate $\tilde{x}_i(k)$ from the optimal value of problem (2).

VI. Conclusion

In this paper, we present a distributed subgradient method for minimizing a sum of convex functions, where each of the component function represents a cost function for an individual agent, known by that agent only. The method involves the agents maintaining estimates of the solution of the global optimization problem and updating them by averaging with neighbors in the network and by taking a subgradient step using their local objective function. Under the assumption that the availability of communication links is represented by a stochastic process, we provide a convergence analysis for this method.

In particular, we consider related estimates $\tilde{x}_i(k)$—the long-run average of the local estimate $x_i(k)$—for each agent i. With diminishing stepsizes, we show that the objective function value (or cost) of the estimates converges with probability 1 to the optimal cost. With a constant stepsize, the objective function value (or cost) of the averaged estimates converges with probability 1 to a neighborhood of the optimal cost.

This paper contributes to a large and growing literature on multi-agent control and optimization. There are many directions in which this research can be extended meaningfully: analyzing this problem with a stochastic process that is not independent and identically distributed over time could allow our subgradient method to be used, for example, in a scenario where the sensors are mobile; relaxing the doubly stochasticity assumption would permit nonsymmetric communication between agents; introducing random message delays would add an important real-world phenomenon to this model; and considering constrained optimization would also add to the applicability of this model.

References

[1] D. P. Bertsekas, A. Nedić, and A. E. Ozdaglar, *Convex Analysis and Optimization*. Cambridge, MA: Athena Scientific, 2003.

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

[3] D. Bertsimas and J. N. Tsitsiklis, *Linear Optimization*. Belmont, MA: Athena Scientific, 1985.

[4] V. D. Blondel, J. M. Hendrickx, A. Olshevsky, and J. N. Tsitsiklis, "Convergence in multiagent coordination, consensus, and flocking," presented at the IEEE CDC, 2005.

[5] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Gossip algorithms: Design, analysis, and applications," presented at the IEEE INFOCOM, 2005.

[6] M. Cao, D. A. Spielman, and A. S. Morse, "A lower bound on convergence of a distributed network consensus algorithm," presented at the IEEE CDC, 2005.

[7] R. Carli, F. Fagnani, A. Speranzon, and S. Zampieri, "Communication constraints in coordinated consensus problems," presented at the IEEE ACC, 2006.

[8] M. Chiang, S. H. Low, A. R. Calderbank, and J. C. Doyle, "Layering as optimization decomposition: A mathematical theory of network architectures," *Proc. IEEE*, vol. 95, pp. 255–312, 2007.

[9] Y. Hatano and M. Mesbahi, "Agreement over random networks," *IEEE Trans. Autom. Control*, vol. 50, no. 11, pp. 1867–1872, Nov. 2005.

[10] A. Jadbabaie, J. Lin, and S. Morse, "Coordination of groups of mobile autonomous agents using nearest neighbor rules," *IEEE Trans. Autom. Control*, vol. 48, no. 9, pp. 988–1001, Sep. 2003.

[11] F. P. Kelly, A. K. Maulloo, and D. K. Tan, "Rate control for communication networks: Shadow prices, proportional fairness, and stability," *J. Oper. Res. Soc.*, vol. 49, pp. 237–252, 1998.

[12] S. Low and D. E. Lapsley, "Optimization flow control, I: Basic algorithm and convergence," *IEEE/ACM Trans. Netw.*, vol. 7, no. 6, pp. 861–874, Dec. 1999.

[13] A. Nedić, A. Olshevsky, A. Ozdaglar, and J. N. Tsitsiklis, "On distributed averaging algorithms and quantization effects," *IEEE Trans. Autom. Control*, vol. 54, no. 1, pp. 2506–2517, Nov. 2009.

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

[15] A. Nedić, A. Ozdaglar, and P. A. Parrilo, "Constrained consensus and optimization in multi-agent networks," *IEEE Trans. Autom. Control*, 2008, LIDS Rep. 2779, to be published.

[16] R. Olfati-Saber and R. M. Murray, "Consensus problems in networks of agents with switching topology and time-delays," *IEEE Trans. Autom. Control*, vol. 49, no. 9, pp. 1520–1533, Sep. 2004.

[17] A. Olshevsky and J. N. Tsitsiklis, "Convergence rates in distributed consensus averaging," presented at the IEEE CDC, 2006.

[18] A. Olshevsky and J. N. Tsitsiklis, "Convergence speed in distributed consensus and averaging," *SIAM J. Control Optim.*, vol. 48, no. 1, pp. 33–55, 2009.

[19] S. S. Ram, A. Nedić, and V. V. Veeravalli, "Stochastic incremental gradient descent for estimation in sensor networks," presented at the Asilomar Conf., 2007.

[20] R. Srikant, *Mathematics of Internet Congestion Control*. New York: Birkhauser, 2004.

[21] A. Tahbaz-Salehi and A. Jadbabaie, "A necessary and sufficient condition for consensus over random networks," *IEEE Trans. Autom. Control*, vol. 53, no. 3, pp. 791–795, Mar. 2008.

[22] J. N. Tsitsiklis, "Problems in Decentralized Decision Making and Computation," Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., Mass. Inst. Technol., Cambridge, MA, 1984.

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

[24] T. Vicsek, A. Czirok, E. Ben-Jacob, I. Cohen, and O. Schochet, "Novel type of phase transitions in a system of self-driven particles," *Phys. Rev. Lett.*, vol. 75, no. 6, pp. 1226–1229, 1995.

[25] C. W. Wu, "Syncronization and convergence of linear dynamics in random directed networks," *IEEE Trans. Autom. Control*, vol. 51, no. 7, pp. 1207–1210, Jul. 2006.

Ilan Lobel received the B.S. degree in electrical engineering from the Pontificia Universidade Catolica of Rio de Janeiro, Brazil, in 2004, and the Ph.D. degree in operations research from the Massachusetts Institute of Technology, Cambridge, in 2009.

In 2010, he joined the New York University's (NYU) Stern School of Business, where he is an Assistant Professor of information, operations, and management sciences. Prior to joining NYU Stern, he was a postdoctoral researcher at Microsoft Research, New England Lab. His research focuses on decision-making in social networks, algorithms for distributed optimization, and dynamic mechanism design.

Asu Ozdaglar (M'95) received the B.S. degree in electrical engineering from the Middle East Technical University, Ankara, Turkey, in 1996, and the M.S. and the Ph.D. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, in 1998 and 2003, respectively.

Since 2003, she has been a member of the faculty of the Electrical Engineering and Computer Science Department, MIT, where she is currently the Class of 1943 Associate Professor. She is affiliated with the Laboratory for Information and Decision Systems and the Operations Research Center at MIT. Her research interests include optimization theory, with emphasis on nonlinear programming and convex analysis, game theory, with applications in communication, social, and economic networks, and distributed optimization and control. She is the co-author of the book entitled *Convex Analysis and Optimization*.

Dr. Ozdaglar is a recipient of a Microsoft fellowship, the MIT GSC Teaching award, the NSF Career award, and the 2008 Donald P. Eckman award of the American Automatic Control Council.