Convergence Analysis of Distributed Subgradient Methods over Random Networks

Ilan Lobel
Operations Research Center
Massachusetts Institute of Technology
Cambridge, MA 02139
lobel@mit.edu

Asuman Ozdaglar
Department of Electrical Engineering
and Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139
asuman@mit.edu

***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 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 convergence results and convergence rate estimates 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 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. 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 [20] (see also Tsitsiklis *et al.* [21], 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.

More specifically, our model involves a set of agents whose goal is to cooperatively minimize a convex cost 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 cost of agent i, known only by this agent. 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 $x_i(k)$.

Under weak assumptions, we prove that for every agent i, the sequence of objective function values of the estimates $\tilde{x}_i(k)$ converges to the optimal cost in expectation when we use a diminishing stepsize rule. We also show that the objective function values converge in expectation to a neighborhood of the optimal cost when we use a constant stepsize rule. We provide rates of convergence and show the trade-off between using larger stepsizes (to obtain faster convergence) and using smaller ones (to get convergence to a better neighborhood).

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

This research was partially supported by the National Science Foundation under Career grant DMI-0545910, and by the DARPA ITMANET program.

recent attention as natural models of cooperative behavior in networked-systems (see Vicsek *et al.* [22], Jadbabaie *et al.* [9], Boyd *et al.* [5], Olfati-Saber and Murray [15], Cao *et al.* [6], Olshevsky and Tsitsiklis [16], [17], and Olshevsky *et al.* [13]). Our work is also related to the *utility maximization* framework for resource allocation in networks (see Kelly *et al.* [10], Low and Lapsley [12], Srikant [18], and Chiang *et al.* [7]). 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, IV and V build the tools that we use to analyze our model: Section III develops some results on the communication networks, Section IV establishes some needed results about products of random matrices and Section V studies properties of the iterates of the subgradient method. Section VI combines lemmas and propositions from the previous sections to prove our convergence results. Section VII concludes the paper.

For space considerations, all proofs are omitted from this paper. They are available in [11].

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:

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

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

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

where the vector $a^i(k) = (a^i_1(k), \ldots, a^i_n(k))'$ is a vector of weights 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 the 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.* [9] and Blondel *et al.* [4].

Let $x^l(k)$ denote the vector of the l^{th} component of all agent estimates at time k, i.e., $x^l(k) = (x^l_1(k), \ldots, x^l_n(k))$ 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 $d^l(k) = (d^l_1(k), \ldots, d^l_n(k))$ is a vector of the l^{th} 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 following section 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 [8] 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 [23] and Tahbaz-Salehi and Jadbabaie [19] 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

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

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 construct random communication events that have the following property: if one such event occurs, then information has been propagated from each agent to every other agent. We establish bounds on the probability of such an event occurring and the 'amount' of information propagated 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 such that $k \geq s \geq 0$,

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

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

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

(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 $k \geq r$.*

Again, all proofs are omitted from this paper and can be found at [11]. 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

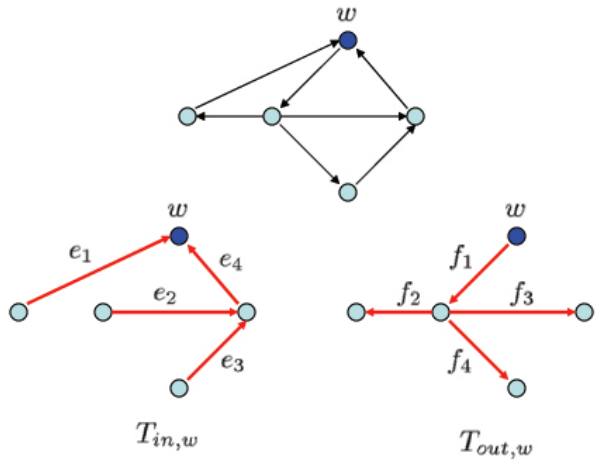

Fig. 1. A strongly connected mean connectivity graph and the two directed spanning trees rooted at node w on this graph. The figure illustrates the labeling of the edges on the in-tree $T_{in,w}$ and the out-tree $T_{out,w}$ according to the procedure described in the text. Note that the edges on all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

$(\mathcal{N}, \tilde{\mathcal{E}})$: an in-tree rooted at w, denoted by $T_{in,w}$ (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 $T_{out,w}$ (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 $T_{in,w}$, 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 $T_{in,r}$, 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 $T_{out,w}$, 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

$$T_{in,w} = \{e_1, e_2, \ldots, e_{n-1}\}, \quad T_{out,w} = \{f_1, f_2, \ldots, f_{n-1}\}, \tag{9}$$

(see Figure 1).

[1]Note that this edge labeling ensures that all edges are labeled in 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 $T_{out,w}$, but it can be seen that this creates a cycle on the tree $T_{out,w}$ – a contradiction.

We next define the probabilistic event that ensures information exchange across the network. Recall that for any edge $e = (j, i)$, the notation A_e denotes the (i, j) entry of the matrix A. Given any time $s \geq 0$, we define the following events for all $l \in \{1, \ldots, n-1\}$,

$$C_l(s) = \{A^\infty \in \Omega^\infty \mid A_{e_l}(s + l - 1) \geq \gamma\}, \tag{10}$$

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

and

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

For all $l = 1, \ldots, n-1$, the event $C_l(s)$ denotes the event that edge $e_l \in T_{in,w}$ is activated at time $s + l - 1$, and the event $D_l(s)$ denotes the event that edge $f_l \in T_{out,w}$ 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 $T_{in,w}$ and $T_{out,w}$ are activated sequentially following time s in the order given in Eq. (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 for all $i, j \in \mathcal{N}$ and $k \geq s + 2(n-1) - 1$,

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

The previous lemma stated 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) *The events $C_l(s)$ and $D_l(s)$ for all $l = 1, \ldots, n-1$ are mutually independent and*

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

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

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., for all $i, j \in \mathcal{N}$,

$$[\Phi(s + 2(n-1) - 1, s)]_{ij} \geq \gamma^{2(n-1)}.$$

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 [20]. 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 $(i,j) \in \{1,...,n\}^2$, $[D_k]_i^j \geq \delta$. 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, i.e., for all $k \geq 1$,*

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

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 $A(k)$, $k = 0, 1, ...$ 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. Therefore, if every pair of agents always uses the same coefficients, i.e., for each $k \geq 0$, $a_{ij}(k) = a_{ji}(k)$ for all $(i,j) \in \{1,...,n\}^2$ with probability 1, then double stochasticity is satisfied.

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 $(i,j) \in \{1,...,n\}^2$ of $\bar{D}$ satisfies $[\bar{D}]_i^j = \frac{1}{n}$. 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 $k \geq 1$,

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

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

$$b(k,s) = \max_{i,j\in\mathcal{N}} \left| [\Phi(k,s)]_i^j - \frac{1}{n} \right| \quad \text{for all } k \geq s. \tag{13}$$

The following lemma states that if t independent events of the form $G(s_i)$, for $i = 1,..,t$, occur between r and k, 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 $G(s_1)$, $i = 1,..,t$.

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

$$b(k,r) \leq 2\left(1+\frac{1}{\gamma^{2(n-1)}}\right)\left(1-\gamma^{2(n-1)}\right)^t. \tag{14}$$

Define the following two constants:

$$C = \left(3+\frac{2}{\gamma^{2(n-1)}}\right)\exp\left\{-\frac{\gamma^{4(n-1)}}{2}\right\} \text{ and} \tag{15}$$

$$\beta = \exp\left\{-\frac{\gamma^{4(n-1)}}{4(n-1)}\right\}. \tag{16}$$

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

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

where β and C are given by Eqs. (15) and (16).

Therefore, we obtain that for all $k \geq s$, $E[b(k,s)]$ decays exponentially in $k-s$. Combined with the results of the next section, this will enable us to bound the solution of the distributed subgradient method.

V. Analysis of the Subgradient Method

In this section, we establish key relations for the iterates of the distributed subgradient method given in Eq. (3), which hold under any stepsize rules. Here, we obtain results in terms of the random variables $b(k,s)$ for $k \geq s$. All results of this section are inequalities between random variables that hold with probability 1.

Using the linearity of the update rule given in Eq. (3) and the definition of the transition matrices [cf. Eq. (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 $k \geq s \geq 0$,

$$\begin{aligned} x_i(k+1) &= \sum_{j=1}^n [\Phi(k,s)]_{ij} x_j(s) \\ &\quad - \sum_{r=s+1}^k \sum_{j=1}^n [\Phi(k,r)]_{ij} \alpha(r-1) d_j(r-1) \\ &\quad - \alpha(k) d_i(k), \end{aligned} \tag{18}$$

[cf. Eq. (7)].

To analyze 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 $k \geq 0$, defined as follows: Let the initial iterate $y(0)$ be given by

$$y(0) = \frac{1}{n}\sum_{j=1}^n x_j(0). \tag{19}$$

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

$$y(k+1) = y(k) - \frac{\alpha(k)}{n}\sum_{j=1}^n d_j(k). \tag{20}$$

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

$$y(k) = \frac{1}{n}\sum_{j=1}^n x_j(0) - \frac{1}{n}\sum_{s=1}^k \alpha(s)\sum_{j=1}^n d_j(s-1). \tag{21}$$

The iterate $y(k)$ represents a 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 (20) 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 $\|s\| \leq L$.*

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

$$\max_{1\leq j\leq n} \|x_j(0)\| \leq L, \tag{22}$$

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 $k \geq 0$. We also consider the (weighted) averaged-vectors $\tilde{x}_i(k)$ and $\tilde{y}(k)$ defined for all $k \geq 0$ as

$$\tilde{x}_i(k) = \frac{1}{\sum_{s=0}^k \alpha(s)} \sum_{t=0}^k \alpha(t) x_i(t) \quad \text{and} \tag{23}$$

$$\tilde{y}(k) = \frac{1}{\sum_{s=0}^k \alpha(s)} \sum_{t=0}^k \alpha(t) y(t), \tag{24}$$

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 the iteration (20), and the sequences $\{x_i(k)\}$ for $i \in \mathcal{N}$ be generated by the iteration (3).

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

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

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

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

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

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

The following lemma provides a relation for $dist^2(y(k), X^*)$, the squared-distance of the iterates $y(k)$ to the optimal solution set X^*, which will be key in establishing objective function value estimates in the subsequent proposition. 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 8: Let the sequence $\{y(k)\}$ be generated by the iteration (20), and the sequences $\{x_i(k)\}$ for $i \in \mathcal{N}$ be generated by the 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 $k \geq 0$. We then have for all $k \geq 0$,

$$dist^2(y(k+1), X^*) \leq dist^2(y(k), X^*) + \frac{2\alpha(k)}{n} \sum_{j=1}^n (\|d_j(k)\| + \|g_j(k)\|)\|y(k) - x_j(k)\| - \frac{2\alpha(k)}{n}[f(y(k)) - f^*] + \frac{\alpha^2(k)}{n^2} \sum_{j=1}^n \|d_j(k)\|^2.$$

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 2: Let Bounded Subgradients assumption hold [cf. Assumption 4]. Let the sequence $\{y(k)\}$ be generated by the iteration (20), and the sequences $\{x_i(k)\}$ for $i \in \mathcal{N}$ be generated by the iteration (3).

(a) *Let $\tilde{y}(k)$ be the averaged vector defined in Eq. (24). An upper bound on the objective function $f(\tilde{y}(k))$ is given by*

$$f(\tilde{y}(k)) \leq f^* + \frac{n}{2\sum_{r=0}^k \alpha(r)} dist^2(y(0), X^*) + \frac{nL^2}{2\sum_{r=0}^k \alpha(r)} \sum_{t=0}^k \alpha^2(t) + \frac{2n^2L^2}{\sum_{r=0}^k \alpha(r)} \sum_{t=0}^k \alpha(t) \sum_{s=0}^{t-1} \alpha(s-1) b(t-1,s) + \frac{4nL^2}{\sum_{r=0}^k \alpha(r)} \sum_{t=0}^k \alpha(t)\alpha(t-1).$$

(b) *Let $\tilde{x}_i(k)$ be the averaged vector defined in Eq. (23). An upper bound on the objective value $f(\tilde{x}_i(k))$ for each i is given by*

$$f(\tilde{x}_i(k)) \leq f^* + \frac{n}{2\sum_{r=0}^k \alpha(r)} dist^2(y(0), X^*) + \frac{nL^2}{2\sum_{r=0}^k \alpha(r)} \sum_{t=0}^k \alpha^2(t) + \frac{3n^2L^2}{\sum_{r=0}^k \alpha(r)} \sum_{t=0}^k \alpha(t) \sum_{s=0}^{t-1} \alpha(s-1) b(t-1,s) + \frac{6nL^2}{\sum_{r=0}^k \alpha(r)} \sum_{t=0}^k \alpha(t)\alpha(t-1).$$

With this proposition, we have completed the set of tools that we need to prove our convergence results.

VI. Convergence Analysis

In this section, we analyze the distributed subgradient method under two stepsize rules: a diminishing stepsize rule, whereby the stepsize sequence $\{\alpha(k)\}$ satisfies $\alpha(k) > 0$ for all k and $\alpha(k) \downarrow 0$, 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.

We begin our analysis with the decreasing stepsize rule.

Assumption 5: (Diminishing Stepsize) *Assume the stepsize $\{\alpha(k)\}$ is a non-increasing sequence and it satisfies*

$$\sum_{k=0}^{\infty} \alpha(k) = \infty \qquad \text{and} \qquad \sum_{k=0}^{\infty} \alpha(k)^2 = A < \infty. \quad (25)$$

The following lemma establishes the long-run behavior of the relevant random variables in the case of diminishing stepsize.

Lemma 9: Let Connectivity, Doubly Stochastic Weights and Diminishing Stepsize assumptions hold [cf. Assumptions 2, 3 and 5]. We have

$$\lim_{k\to\infty} E\left[\frac{1}{\sum_{r=0}^{k}\alpha(r)}\sum_{t=0}^{k}\sum_{s=0}^{t-1}\alpha(t)\alpha(s-1)b(t-1,s)\right] = 0 \quad (26)$$

and

$$\liminf_{k\to\infty} \frac{1}{\sum_{r=0}^{k}\alpha(r)}\sum_{t=0}^{k}\sum_{s=0}^{t-1}\alpha(t)\alpha(s-1)b(t-1,s) = 0 \quad (27)$$

with probability 1.

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

Theorem 1: Let Connectivity, Doubly Stochastic Weights, Bounded Subgradients and Diminishing Stepsize assumptions hold [cf. Assumptions 2, 3, 4, 5]. Let the sequences $\{x_i(k)\}$ for $i \in \mathcal{N}$ be generated by the iteration (3), and $\tilde{x}_i(k)$ be the averaged vector defined in Eq. (23). For all $i \in \mathcal{N}$, we have

$$\lim_{k\to\infty} E[f(\tilde{x}_i(k))] = f^* \quad \text{and}$$

$$\liminf_{k\to\infty} f(\tilde{x}_i(k)) = f^*$$

with probability 1.

Our last result shows that with a constant stepsize rule, we can bound the quality of the expected solution of the distributed subgradient method.

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 $k \geq 0$,

$$E[|f(\tilde{x}_i(k)) - f^*|] \leq \frac{n\ dist^2(y(0), X^*)}{2\alpha(k+1)} + \frac{3Cn^2L^2\alpha}{1-\beta} + \frac{13nL^2\alpha}{2}. \quad (28)$$

Eq. (28) indicates the trade-offs available when choosing the constant parameter α of the subgradient method. If α is too small and the distance from the starting average to the optimal solution – $dist^2(y(0), X^*)$ – is large, then the algorithm will take many iterations to converge. If, on the other hand, α is too large, the region where we expect the subgradient method to converge to becomes larger, thus decreasing the quality of the solution.

VII. Conclusions

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 averaged estimates converges in expectation to the optimal cost. We also show that the limit inferior of the cost sequence converges to the optimal cost with probability one. With a constant stepsize, the objective function value (or cost) of the averaged estimates converges in expectation to a neighborhood of the optimal cost. In this case, we highlight the trade-off between choosing a larger α and obtaining faster convergence and selecting a smaller α in order to obtain convergence to a better neighborhood. We show explicitly the rates of convergence as a function of the parameters of the model and the input parameters of the algorithm.

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 non-symmetric 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*, Athena Scientific, Cambridge, Massachusetts, 2003.

[2] D.P. Bertsekas and J.N. Tsitsiklis, *Parallel and distributed computation: Numerical methods*, Athena Scientific, Belmont, MA, 1997.

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

[4] V.D. Blondel, J.M. Hendrickx, A. Olshevsky, and J.N. Tsitsiklis, *Convergence in multiagent coordination, consensus, and flocking*, Proceedings of IEEE CDC, 2005.

[5] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, *Gossip algorithms: Design, analysis, and applications*, Proceedings of IEEE INFOCOM, 2005.

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

[7] M. Chiang, S.H. Low, A.R. Calderbank, and J.C. Doyle, *Layering as optimization decomposition: A mathematical theory of network architectures*, Proceedings of the IEEE **95** (2007), no. 1, 255–312.

[8] Y. Hatano and M. Mesbahi, *Agreement over random networks*, IEEE Transactions on Automatic Control **50** (2005), no. 11, 1867–1872.

[9] A. Jadbabaie, J. Lin, and S. Morse, *Coordination of groups of mobile autonomous agents using nearest neighbor rules*, IEEE Transactions on Automatic Control **48** (2003), no. 6, 988–1001.

[10] F.P. Kelly, A.K. Maulloo, and D.K. Tan, *Rate control for communication networks: shadow prices, proportional fairness, and stability*, Journal of the Operational Research Society **49** (1998), 237–252.

[11] I. Lobel and A. Ozdaglar, *Distributed subgradient methods over random networks*, Preprint, 2008.

[12] S. Low and D.E. Lapsley, *Optimization flow control, I: Basic algorithm and convergence*, IEEE/ACM Transactions on Networking **7** (1999), no. 6, 861–874.

[13] A. Nedić, A. Olshevsky, A. Ozdaglar, and J.N. Tsitsiklis, *On distributed averaging algorithms and quantization effects*, LIDS report 2778, 2008.

[14] A. Nedić and A. Ozdaglar, *Distributed subgradient methods for multi-agent optimization*, to appear in IEEE Transactions on Automatic Control, 2008.

[15] R. Olfati-Saber and R.M. Murray, *Consensus problems in networks of agents with switching topology and time-delays*, IEEE Transactions on Automatic Control **49** (2004), no. 9, 1520–1533.

[16] A. Olshevsky and J.N. Tsitsiklis, *Convergence rates in distributed consensus averaging*, Proceedings of IEEE CDC, 2006.

[17] ———, *Convergence speed in distributed consensus and averaging*, to appear in SIAM Journal on Control and Optimization, 2008.

[18] R. Srikant, *Mathematics of Internet congestion control*, Birkhauser, 2004.

[19] A. Tahbaz-Salehi and A. Jadbabaie, *A necessary and sufficient condition for consensus over random networks*, to appear in IEEE Transactions on Automatic Control, 2008.

[20] J.N. Tsitsiklis, *Problems in decentralized decision making and computation*, Ph.D. thesis, Dept. of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 1984.

[21] J.N. Tsitsiklis, D.P. Bertsekas, and M. Athans, *Distributed asynchronous deterministic and stochastic gradient optimization algorithms*, IEEE Transactions on Automatic Control **31** (1986), no. 9, 803–812.

[22] T. Vicsek, A. Czirok, E. Ben-Jacob, I. Cohen, and Schochet O., *Novel type of phase transitions in a system of self-driven particles*, Physical Review Letters **75** (1995), no. 6, 1226–1229.

[23] C.W. Wu, *Syncronization and convergence of linear dynamics in random directed networks*, IEEE Transactions on Automatic Control **51** (2006), no. 7, 1207–1210.