52nd IEEE Conference on Decision and Control
December 10-13, 2013. Florence, Italy

Distributed optimization over time-varying directed graphs

Angelia Nedić and Alex Olshevsky

***Abstract*— We consider distributed optimization by a collection of nodes, each having access to its own convex function, whose collective goal is to minimize the sum of the functions. The communications between nodes are described by a time-varying sequence of *directed* graphs, which is uniformly strongly connected. For such communications, assuming that every node knows its out-degree, we develop a broadcast-based algorithm, termed the *subgradient-push*, which steers every node to an optimal value under a standard assumption of subgradient boundedness. The subgradient-push requires no knowledge of either the number of agents or the graph sequence to implement. Our analysis shows that the subgradient-push algorithm converges at a rate of $O\left(\ln t/\sqrt{t}\right)$, where the constant depends on the initial values at the nodes, the subgradient norms, and, more interestingly, on both the consensus speed and the imbalances of influence among the nodes.**

I. Introduction

We consider the problem of distributed convex optimization by a network of nodes when knowledge of the objective function is scattered throughout the network and unavailable at any singe location. There has been much recent interest in multi-agent optimization problems of this type that arise whenever a large collections of nodes - which may be processors, nodes of a sensor network, vehicles, or UAVs - desire to collectively optimize a global objective by means of local actions taken by each node without any centralized coordination.

Specifically, we will study the problem of optimizing a sum of n convex functions by a network of n nodes when each function is known to only a single node. This problem frequently arises when control and signal processing protocols need to be implemented in sensor networks. For example, the problems including robust statistical inference [15], formation control, non-autonomous power control [18], distributed message routing [14], and spectrum access coordination [8], can be reduced to variations of this problem. We will be focusing on the case when communication between nodes is *directed* and *time-varying*.

Distributed optimization of a sum of convex functions has received a surge of interest in recent years [13], [15], [6], [11], [9], [10], [19], [12], [2], [16], [4]. There is now a considerable theory justifying the use of distributed subgradient methods in this setting, and their performance limitations and convergence times are well-understood. Moreover, distributed subgradient methods have been used to propose new solutions for a number of problems in distributed control and sensor networks [18], [14], [8]. However, the works cited above assumed communications among nodes are either fixed or undirected.

Our paper is the first to demonstrate a working subgradient protocol in the setting of directed time-varying communications. We develop a broadcast-based protocol, termed the *subgradient-push*, which steers every node to an optimal value under a standard assumption of subgradient boundedness. The subgradient-push requires each node to know its out-degree at all times, but beyond this it needs no knowledge of the graph sequence of even of the number of agents to implement. Our results show that it converges at a rate of $O\left(\ln t/\sqrt{t}\right)$, where the constant depends, among other factors, on the consensus speed of the corresponding directed graph sequence and a measure of the imbalance of influence among the nodes.

Our work is closest to the recent papers [20], [21], [5]. The papers [20], [21] proved the convergence of a subgradient algorithm in a directed but fixed topology; implementation of the protocol appears to require knowledge of the graph or of the number of agents. By contrast, our results work in time-varying networks and are fully distributed, requiring no knowledge of either the graph sequence or the number of agents. The paper [5] shows the convergence of a distributed optimization protocol in continuous time, also for directed but fixed graphs; moreover, an additional assumption is made in [5] that the graph is "balanced."

All the prior work in distributed optimization, except for [20], [21] requires time-varying communications with some form of balancedness, often reflected in a requirement of having a sequence of doubly stochastic matrices that are commensurate with the sequence of underlying communication graphs. In contrast, our proposed method removes the need for the doubly stochastic matrices. The proposed distributed optimization model is motivated by applications that are characterized by time-varying directed communications such as those arising in a mobile sensor network communication where the links between nodes will come and go as nodes move in and out of line-of-sight or broadcast range of each other. Moreover, if different nodes are capable of broadcasting messages at different power levels, then communication links connecting the nodes will necessarily be unidirectional.

The remainder of this paper is organized as follows. We begin in Section II where we describe the problem of interest, outline the subgradient-push algorithm, and state the main convergence results. Section III is devoted to the proof of a key lemma, namely the convergence rate result for a perturbed version of the so-called push-sum protocol; this

The authors are with the Department of Industrial and Enterprise Systems Engineering, University of Illinois at Urbana-Champaign, 104 S. Mathews Avenue, Urbana IL, 61801, Emails: {angelia,aolshev2}@illinois.edu. A. Nedić gratefully acknowledges the support by the NSF under grants CMMI 07-42538 and CCF 11-11342.

978-1-4673-5716-6/13/$31.00 ©2013 IEEE

lemma is then used in the subsequent proofs of convergence and convergence rate for the subgradient-push in Section IV.

Notation: We will apply boldface to distinguish between the vectors in $\mathbb{R}^d$ and scalars associated with different nodes. For example, the vector $\mathbf{x}_i(t)$ is in boldface to identify a vector for node i, while the scalar $y_i(t)$ is not - which identifies a scalar value for node i. Additionally, for a vector $\mathbf{x}_i$ that has a subscript i identifying a node index, we will use $[\mathbf{x}]_j$ to denote its j'th entry. The vectors such as $y(t) \in \mathbb{R}^n$ obtained by stacking scalar values $y_i(t)$ associated with the nodes is not bolded. For a matrix A, we will use $[A]_{ij}$ to denote the i,j'th entry of A. The vectors are seen as column vectors unless otherwise explicitely stated. We use $\mathbf{1}$ to denote the vector of ones, and $\|y\|$ for the Euclidean norm of a vector y.

II. Problem, Algorithm and Main Results

We consider a network of n nodes whose goal is to minimize the function

$$F(\mathbf{z}) = \sum_{i=1}^{n} f_i(\mathbf{z})$$

where only node i knows the convex function $f_i(\mathbf{z}) : \mathbb{R}^d \to \mathbb{R}$. Under the assumption that the set of optimal solutions $Z^* = \arg\min_{\mathbf{z}\in\mathbb{R}^d} F(\mathbf{z})$ is nonempty, we would like to design a protocol in which all agents will maintain variables $\mathbf{z}_i(t)$ such that all the $\mathbf{z}_i(t)$ converge to the same point in Z^*.

We will assume that, at each time t, *node i can only send messages to its out-neighbors in some directed graph $G(t)$*. Naturally, the graph $G(t)$ will have vertex set $\{1,\ldots,n\}$, and we will use $E(t)$ to denote its edge set. Also, naturally, the sequence $\{G(t)\}$ should posses some good long-term connectivity properties. A standard assumption, which we will be making, is that the sequence $\{G(t)\}$ is uniformly strongly connected (or, as it is sometimes called, B-strongly-connected), namely, that there exists some ineger $B > 0$ (possibly unknown to the nodes) such that the graph with edge set

$$E_B(k) = \bigcup_{i=kB}^{(k+1)B-1} E(i)$$

is strongly connected for every $k \geq 0$. This is a typical assumption for many results in multi-agent control: it is considerably weaker than requiring each $G(t)$ be connected for it allows the edges necessary for connectivity to appear over a long time period and in arbitrary order; however, it is still strong enough to derive bounds on the speed of information propagation from one part of the network to another.

Finally, we introduce the notation $N_i^{\rm in}(t)$ and $N_i^{\rm out}(t)$ for the in- and out-neighborhoods of node i, respectively, at time t. We will allow these neighborhoods to include the node i itself[1]; formally, we have

$$\begin{aligned} N_i^{\rm in}(t) &= \{j \mid (j,i) \in E(t)\} \cup \{i\}, \\ N_i^{\rm out}(t) &= \{j \mid (i,j) \in E(t)\} \cup \{i\}, \end{aligned}$$

and $d_i(t)$ for the out-degree of node i, i.e.,

$$d_i(t) = |N_i^{\rm out}(t)|.$$

Crucially, we will be assuming that *every node i knows its out-degree $d_i(t)$ at every time t*.

Our main result in is a protocol which successfully accomplishes the task of distributed minimization of $F(\mathbf{z})$ under the assumptions we have laid out above. Our scheme is a combination of subgradient descent and the so-called *push-sum* protocol, recently studied in the papers [1], [3], [7]. We will refer to our protocol as the *subgradient-push* method.

A. The subgradient-push method

Every node i will maintain auxiliary vector variables $\mathbf{x}_i(t), \mathbf{w}_i(t)$ in $\mathbb{R}^d$, as well as an auxiliary scalar variable $y_i(t)$, initialized as $y_i(0) = 1$ for all i. These quantities will be updated by the nodes according to the rules,

$$\begin{aligned} \mathbf{w}_i(t+1) &= \sum_{j\in N_i^{\rm in}(t)} \frac{\mathbf{x}_j(t)}{d_j(t)}, \\ y_i(t+1) &= \sum_{j\in N_i^{\rm in}(t)} \frac{y_j(t)}{d_j(t)}, \\ \mathbf{z}_i(t+1) &= \frac{\mathbf{w}_i(t+1)}{y_i(t+1)}, \\ \mathbf{x}_i(t+1) &= \mathbf{w}_i(t+1) - \alpha(t+1)\mathbf{g}_i(t+1), \end{aligned} \quad (1)$$

where $\mathbf{g}_i(t+1)$ is a subgradient of the function f_i at $\mathbf{z}_i(t+1)$. The method is initiated with $\mathbf{w}_i(0) = \mathbf{z}_i(0) = \mathbf{1}$ and $y_i(0) = 1$ for all i. The stepsize $\alpha(t+1) > 0$ satisfies the following decay conditions for all $t > s \geq 0$

$$\sum_{t=1}^{\infty} \alpha(t) = \infty, \qquad \sum_{t=1}^{\infty} \alpha^2(t) < \infty, \qquad \alpha(t) \leq \alpha(s) \quad (2)$$

We note that the above equations have simple broadcast-based implementation: each node i broadcasts the quantities $\mathbf{x}_i(t)/d_i(t), y_i(t)/d_i(t)$ to all of the nodes in its out-neighborhood[2], which simply sum all the messages they receive to obtain $\mathbf{w}_i(t+1)$ and $y_i(t+1)$. The update equations for $\mathbf{z}_i(t+1), \mathbf{x}_i(t+1)$ can then be executed without any further communications between nodes during step t.

Without the subgradient term in the final equation, our protocol would be a version of the push-sum protocol [7] for average computation studied recently in [1], [3]. For intuition on the precise form of these equations, we refer the reader to these three papers; roughly speaking, the somewhat involved form of the updates is intended to ensure that every node receives an equal weighting after all the linear

[1] Alternatively, one may define these neighborhoods in a standard way of the graph theory, but require that each graph in the sequence $\{G(t)\}$ has a self-loop at every node.

[2] We note that we make use here of the assumption that node i knows its out-degree $d_i(t)$.

combinations and ratios have been taken. In this case, the vectors $\mathbf{z}_i(t+1)$ converge to some common point, i.e., a consensus is achieved. The inclusion of the subgradient terms in the updates of $\mathbf{x}_i(t+1)$ is intended to steer the consensus point towards the optimal set Z^*, while the push-sum updates steer the vectors $\mathbf{z}_i(t+1)$ towards each other. Our main results, which we describe in the next section, demonstrate that this scheme succeeds in steering all vectors $\mathbf{z}_i(t+1)$ towards the same point in the solution set Z^*.

B. Our results

Our first theorem demonstrates the correctness of the subgradient-push method for an arbitrary stepsize $\alpha(t)$ satisfying Eq. (2); this holds under the assumptions we have laid out above, as well as an additional technical assumption on the boundedness of the subgradients.

Theorem 1. *Suppose that:*

(a) *The graph sequence $\{G(t)\}$ is uniformly strongly connected with a self-loop at every node.*

(b) *Each function $f_i(\mathbf{z})$ is convex and the set $Z^* = \arg\min_{\mathbf{z}\in\mathbb{R}^d} F(\mathbf{z})$ is nonempty.*

(c) *The subgradients of each $f_i(\mathbf{z})$ are uniformly bounded, i.e., there exists $L_i < \infty$ such that*

$$\|\mathbf{g}_i\|_2 \leq L_i.$$

for all subgradients $\mathbf{g}_i$ of $f_i(\mathbf{z})$ at all points $\mathbf{z} \in \mathbb{R}^d$

Then, the distributed subgradient-push method of Eq. (1) *with the stepsize satisfying the conditions in Eq.* (2) *has the following property*

$$\lim_{t\to\infty} \mathbf{z}_i(t) = \mathbf{z}^* \qquad \text{for all } i \text{ and for some } \mathbf{z}^* \in Z^*.$$

Our second theorem makes explicit the rate at which the objective function converges to its optimal value. As standard with subgradient methods, we will make two tweaks in order to get a convergence rate result: (i) we take a stepsize which decays as $\alpha(t) = 1/\sqrt{t}$ (stepsizes which decay at faster rates usually produce inferior convergence rates), and (ii) each node i will maintain a convex combination of the values $\mathbf{z}_i(1), \mathbf{z}_i(2), \ldots$ for which the convergence rate will be obtained. We then demonstrate that the subgradient-push converges at a rate of $O(\ln t/\sqrt{t})$; this is formally stated in the following theorem. The theorem makes use of the matrix $A(t)$ that captures the weights used in the construction of $\mathbf{w}_i(t+1)$ and $y_i(t+1)$ in Eq. (1), which are defined by

$$A_{ij}(t) = \begin{cases} 1/d_j(t) & \text{whenever } j \in N_i^{\text{in}}(t), \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Theorem 2. *Suppose all the assumptions of Theorem 1 hold and, additionally, $\alpha(t) = 1/\sqrt{t}$ for $t \geq 1$. Moreover, suppose that every node i maintains the variable $\widetilde{\mathbf{z}}_i(t) \in \mathbb{R}^d$ initialized at time $t = 1$ to $\widetilde{\mathbf{z}}_i(1) = \mathbf{z}_i(1)$ and updated as*

$$\widetilde{\mathbf{z}}_i(t+1) = \frac{\alpha(t+1)\mathbf{z}_i(t+1) + S(t)\widetilde{\mathbf{z}}_i(t)}{S(t+1)},$$

where $S(t) = \sum_{s=0}^{t-1} \alpha(s+1)$. Then, we have that for all $t \geq 1$, $i = 1, \ldots, n$, and any $\mathbf{z}^* \in Z^*$,

$$\begin{aligned} F(\widetilde{\mathbf{z}}(t)) - F(\mathbf{z}^*) \quad \leq \quad & \frac{n}{2}\frac{\|\bar{\mathbf{x}}(0) - \mathbf{z}^*\|_1}{\sqrt{t}} + \frac{n}{2}\frac{\left(\sum_{i=1}^n L_i\right)^2}{4}\frac{(1+\ln t)}{\sqrt{t}} \\ & + \frac{16}{\delta(1-\lambda)}\left(\sum_{i=1}^n L_i\right)\frac{\sum_{j=1}^n \|\mathbf{x}_j(0)\|_1}{\sqrt{t}} \\ & + \frac{16}{\delta(1-\lambda)}\left(\sum_{i=1}^n L_i^2\right)\frac{(1+\ln t)}{\sqrt{t}} \end{aligned}$$

where

$$\bar{\mathbf{x}}(0) = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i(0),$$

and the scalars λ and δ are functions of the graph sequence $G(1), G(2), \ldots$, which have the following properties:

(a) *For any B-connected graph sequence with a self-loop at every node,*

$$\begin{aligned} \delta \quad &\geq \quad \frac{1}{n^{nB}}, \\ \lambda \quad &\leq \quad \left(1 - \frac{1}{n^{nB}}\right)^{1/(nB)}. \end{aligned}$$

(b) *If each of the graphs $G(t)$ is regular*[3]*, then*

$$\begin{aligned} \delta \quad &= \quad 1 \\ \lambda \quad &\leq \quad \min\left\{\left(1 - \frac{1}{4n^3}\right)^{1/B}, \max_{t\geq 1}\sqrt{\sigma_2(A(t))}\right\} \end{aligned}$$

where $A(t)$ is defined by Eq. (3) *and $\sigma_2(A)$ is the second-largest singular value of a matrix A.*

Several features of this theorem are expected: it is standard for a distributed subgradient method to converge at a rate of $O(\ln t/\sqrt{t})$ with the constant depending on the subgradient-norm upper bounds L_i, as well as on the initial conditions $\mathbf{x}_i(0)$ [17], [4]. Moreover, it is also standard for the analysis of these method to involve λ, which is a measure of the connectivity of the directed sequence $G(1), G(2), \ldots$; namely, the closeness of λ to 1 measures the speed at which a consensus process on the graph sequence $\{G(t)\}$ converges.

However, our bounds also include the parameter δ, which, as we will later see, is a measure of the imbalance of influences among the nodes. Time-varying directed regular networks are uniform in influence and will have $\delta = 1$, so that δ will disappear from the bounds entirely; however, networks which are, in a sense to be specified, non-uniform will suffer a corresponding blow-up in the convergence time of the subgradient-push algorithm.

Moreover, we note that while the term $1/(\delta(1-\lambda))$ appearing in our bounds is bounded only exponentially as n^{2nB} in the worst case, it need not be this large for every graph sequence; indeed, part (b) of Theorem 2 shows that for a class of time-varying directed graphs, $1/(\delta(1-\lambda))$ scales polynomially in n. Our work therefore motivates the question of obtaining effective bounds on consensus speed

[3]The graph $G(t)$ is regular if there exists some $d(t)$ such that every out-degree and every in-degree of a node in $G(t)$ equals $d(t)$.

and imbalance of the influence in sequences of directed graphs. Finally, we remark that previous research [13], [17], [4] has studied the case when the matrices $A(t)$ (defined in the statement of Theorem 2) are doubly stochastic; this occurs when the directed graph sequence $\{G(t)\}$ is regular, and in that case our polynomial bounds essentially match previously known results.

III. Perturbed Push-Sum Protocol

This section is dedicated to the analysis a perturbed version of the so-called push-sum protocol, originally introduced in the groundbreaking work [7] and recently analyzed in time-varying directed graphs in [1], [3]. The push-sum is a protocol for node interaction in directed topologies which allows nodes to compute averages and other aggregates in spite of the one-way nature of the communication links.

The original results of [7], [1], [3] demonstrate the convergence of the push-sum protocol. Here we will prove a generalization of this fact by showing that the protocol remains convergent even if the state of the nodes is perturbed at each step, as long as the size of the perturbations decays to zero. We will later use this result in the proof Theorems 1 and 2; because it has has a self-contained interpretation and analysis, we sequester it to this section.

We begin with a statement of the perturbed push-sum update rule. Every node i maintains scalar variables $x_i(t), y_i(t), z_i(t), w_i(t)$ where we assume $y_i(0) = 1$ for all $i = 1, \ldots, n$. These variables are updated as follows:

$$\begin{aligned} w_i(t+1) &= \sum_{j \in N_i^{\text{in}}(t)} \frac{x_j(t)}{d_j(t)}, \\ y_i(t+1) &= \sum_{j \in N_i^{\text{in}}(t)} \frac{y_j(t)}{d_j(t)}, \\ z_i(t+1) &= \frac{w_i(t+1)}{y_i(t+1)}, \\ x_i(t+1) &= w_i(t+1) + \epsilon_i(t+1), \end{aligned} \tag{4}$$

where $\epsilon_i(t)$ is a perturbation at every step, perhaps adversarially chosen. We assume that $N_i^{\text{in}}(t)$ is the in-neighborhood of node i in a directed graph $G(t)$ and $d_j(t)$ is the out-degree of node j, as previously defined in Section II.

We note that without the perturbation term $\epsilon_i(t)$, the method in Eq. (4) reduces to the push-sum protocol. Moreover, our proposed subgradient-push method of Eq. (1) is simply Eq. (4) with a specific form for these perturbation vectors $\epsilon_i(t)$.

The precise form of the push-sum equations of Eq. (4) is a little involved. These dynamics were introduced for the purpose of average computation (in the case when all the perturbations $\epsilon_i(t)$ are zero) and have a simple motivating intuition. The push-sum is a variation of a consensus-like protocol wherein every node updates its values by taking linear combinations of the values of its neighbors; in such schemes, some nodes are bound to be more influential than others (meaning that other nodes end up placing higher coefficients on them), for example by virtue of being more centrally placed. The dynamics of the push-sum are designed around the ratio $z_i(t) = w_i(t)/y_i(t)$ in which these imbalances of influence are meant to be cancelled so that each $z_i(t)$ converges to $(1/n)\sum_{i=1}^n x_i(0)$. We refer the reader to [7], [3], [1] for more details.

We may rewrite the perturbed push-sum equations in more compact form. Using the definition of $A(t)$ from Eq. (3), the relations in Eq. (4) assume the following form:

$$\begin{aligned} w(t+1) &= A(t)x(t), \\ y(t+1) &= A(t)y(t), \\ z_i(t+1) &= \frac{w_i(t+1)}{y_i(t+1)}, \\ x(t+1) &= w(t+1) + \epsilon(t+1), \end{aligned} \tag{5}$$

where $\epsilon(t) = (\epsilon_1(t), \ldots, \epsilon_n(t))'$. Observe that each of the matrices $A(t)$ is column-stochastic but not necessarily row-stochastic.

We will be concerned here with demonstrating a convergence result and a convergence rate for the updates given in Eq. (4), or equivalently, in Eq. (5). Specifically, the bulk of this section is dedicated to proving the following lemma.

Lemma 3. *Consider the sequences $\{z_i(t)\}$, $i = 1, \ldots, n$, generated by the method in Eq. (4). Assuming that the graph sequence $\{G(t)\}$ is uniformly strongly connected, the following statements hold:*

(a) *There exists some $\delta > 0$ and $\lambda \in (0, 1)$ such that for all $t \geq 1$ we have*

$$\left| z_i(t+1) - \frac{\mathbf{1}'x(t)}{n} \right| \leq \frac{8}{\delta}\left(\lambda^t \|x(0)\|_1 + \sum_{s=1}^{t} \lambda^{t-s}\|\epsilon(s)\|_1\right),$$

Moreover, we may choose δ, λ satisfying

$$\delta \geq \frac{1}{n^{nB}}, \quad \lambda \leq \left(1 - \frac{1}{n^{nB}}\right)^{1/B}.$$

If in addition each of the matrices $A(t)$ is doubly stochastic, then

$$\delta = 1, \qquad \lambda \leq \left\{\left(1 - \frac{1}{4n^3}\right)^{1/B}, \max_{t \geq 1} \sqrt{\sigma_2(A(t))}\right\}.$$

(b) *If $\lim_{t\to 0} \epsilon_i(t) = 0$ for all $i = 1, \ldots, n$, then*

$$\lim_{t \to 0} \left| z_i(t+1) - \frac{\mathbf{1}'x(t)}{n} \right| = 0.$$

(c) *If $\{\alpha(t)\}$ is a non-increasing positive scalar sequence such that $\sum_{t=1}^{\infty} \alpha(t)|\epsilon_i(t)| < \infty$ for all i, then*

$$\sum_{t=0}^{\infty} \alpha(t+1) \left| z_i(t+1) - \frac{\mathbf{1}'x(t)}{n} \right| < \infty \qquad \text{for all } i.$$

For part (b) of Lemma 3, observe that each of the matrices $A(t)$ is doubly stochastic if each of the graphs $G(t)$ is regular. Furthermore, we observe that if $\epsilon_i(t) = 0$, this lemma implies that the push-sum method converges at a geometric rate; moreover, it is easy to see that $\mathbf{1}'x(t)/n =$

$\mathbf{1}'x(0)/n$ and therefore $z_i(t) \to \mathbf{1}'x(0)/n$, so that the push-sum protocol successfully computes the average. In the more general case when the perturbations are nonzero, the lemma states that if these perturbations decay to zero, then the push-sum method still converges. Of course, it will no longer be true in this case that the convergence is necessarily on the average of the initial values.

We conclude this section by remarking that Lemma 3 holds even if $x_i(t)$ (and, by extension, $z_i(t)$) is a d-dimensional vector, by applying the results to each coordinate component of the space.

IV. Convergence Results for Subgradient-Push Method

We turn now to the proofs of our main results, Theorems 1 and 2. Our arguments will crucially rely on the convergence results for the perturbed push-sum method we have established in the previous section.

We give a brief, informal summary of the main ideas behind our argument. The convergence result for the perturbed push-sum method of the previous section implies that, under the appropriate assumptions, the entries of $\mathbf{z}_i(t)$ get close to each other over time, and consequently $\mathbf{z}_i(t)$ approaches a multiple of the all-ones vector. Thus every node takes a subgradient of its own function f_i at nearly the same point; over time, these subgradients are averaged by the push-sum-like updates of our method, and the subgradient push approximates the ordinary subgradient algorithm applied to the average function $\frac{1}{n}\sum_{j=1}^n f_j$.

Lemma 4. *Under the same assumptions as in Theorem 1, we have for all $\mathbf{v} \in \mathbb{R}^d$ and $t \geq 0$:*

$$\begin{aligned}\|\bar{\mathbf{x}}(t+1)-\mathbf{v}\|^2 \quad &\leq \quad \|\bar{\mathbf{x}}(t)-\mathbf{v}\|^2 - \frac{2\alpha(t+1)}{n}\left(F(\bar{\mathbf{x}}(t))-F(\mathbf{v})\right)\\ &\quad +\frac{4\alpha(t+1)}{n}\sum_{i=1}^n L_i\|\mathbf{z}_i(t+1)-\bar{\mathbf{x}}(t)\|\\ &\quad +\alpha^2(t+1)\frac{(\mathbf{1}'L_F)^2}{n^2}.\end{aligned}$$

Proof. Let us define $\widetilde{x}_\ell(t)$ to be the vector in $\mathbb{R}^n$ which stacks up the ℓ'th entries of all the vectors $\mathbf{x}_i(t)$: formally, we define $\widetilde{x}_\ell(t)$ to be the vector whose j'th entry is the ℓ'th entry of $\mathbf{x}_j(t)$. Similarly, we define $\widetilde{g}_\ell(t)$ to be the vector stacking up the ℓ'th entries of the vectors $\mathbf{g}_i(t)$: the j'th entry of $\widetilde{g}_\ell(t)$ is the ℓ'th entry of $\mathbf{g}_j(t)$.

It is easy to see from the definition of the subgradient-push in Eq. (1) that

$$\widetilde{x}_\ell(t+1) = A(t)\widetilde{x}_\ell(t) - \alpha(t+1)\widetilde{g}_\ell(t+1) \qquad \text{for } \ell = 1,\ldots,d.$$

Since $A(t)$ is a column-stochastic matrix, this implies for all $\ell = 1,\ldots,d$,

$$\frac{1}{n}\sum_{j=1}^n [\widetilde{x}_\ell(t+1)]_j = \frac{1}{n}\sum_{j=1}^n [\widetilde{x}_\ell(t)]_j - \frac{\alpha(t+1)}{n}\sum_{j=1}^n [\widetilde{g}_\ell(t+1)]_j.$$

Since the ℓ'th entry of $\bar{\mathbf{x}}(t+1)$ is exactly the left-hand side above, we can conclude that

$$\bar{\mathbf{x}}(t+1) = \bar{\mathbf{x}}(t) - \frac{\alpha(t+1)}{n}\sum_{j=1}^n \mathbf{g}_j(t+1). \tag{6}$$

Now let $\mathbf{v} \in \mathbb{R}^d$ be an arbitrary vector. From relation (6) it follows that for all $t \geq 0$,

$$\begin{aligned}\|\bar{\mathbf{x}}(t+1)-\mathbf{v}\|^2 = \|\bar{\mathbf{x}}(t)-\mathbf{v}\|^2 - \frac{2\alpha(t+1)}{n}\sum_{i=1}^n \mathbf{g}_i'(t+1)(\bar{\mathbf{x}}(t)-\mathbf{v})\\ +\frac{\alpha^2(t+1)}{n^2}\left\|\sum_{i=1}^n \mathbf{g}_i(t+1)\right\|^2.\end{aligned}$$

Since the subgradient norms of each f_i are uniformly bounded by L_i, it further follows that for all $t \geq 0$,

$$\begin{aligned}\|\bar{\mathbf{x}}(t+1)-\mathbf{v}\|^2 \leq \|\bar{\mathbf{x}}(t)-\mathbf{v}\|^2 - \frac{2\alpha(t+1)}{n}\sum_{i=1}^n \mathbf{g}_i'(t+1)(\bar{\mathbf{x}}(t)-\mathbf{v})\\ +\alpha^2(t+1)\frac{(\mathbf{1}'L_F)^2}{n^2},\end{aligned} \tag{7}$$

where $L_F = (L_1,\ldots,L_n)'$.

We next consider the cross-term $\sum_{i=1}^n \mathbf{g}_i'(t+1)(\bar{\mathbf{x}}(t)-\mathbf{v})$ in (7). For this term, we write

$$\begin{aligned}\sum_{i=1}^n \mathbf{g}_i'(t+1)(\bar{\mathbf{x}}(t)-\mathbf{v}) \quad &= \quad \sum_{i=1}^n \mathbf{g}_i'(t+1)((\bar{\mathbf{x}}(t)-\mathbf{z}_i(t+1))\\ &\quad +(\mathbf{z}_i(t+1)-\mathbf{v})).\end{aligned} \tag{8}$$

Using the subgradient boundedness and Cauchy-Schwarz, we can lower bound the first term $\mathbf{g}_i'(t+1)(\bar{\mathbf{x}}(t)-\mathbf{z}_i(t+1))$ as

$$\mathbf{g}_i'(t+1)(\bar{\mathbf{x}}(t)-\mathbf{z}_i(t+1)) \geq -L_i\|\bar{\mathbf{x}}(t)-\mathbf{z}_i(t+1)\|.$$

As for the second term $\mathbf{g}_i'(t+1)(\mathbf{z}_i(t+1)-\mathbf{v})$, we can use the fact that $\mathbf{g}_i'(t+1)$ is the subgradient of $f_i(\theta)$ at $\theta = \mathbf{z}_i(t+1)$:

$$\mathbf{g}_i(t+1)(\mathbf{z}_i(t+1)-\mathbf{v}) \geq f_i(\mathbf{z}_i(t+1)) - f_i(\mathbf{v}),$$

from which, by adding and subtracting $\bar{\mathbf{x}}(t)$ and using the Lipschitz continuity of f_i (implied by the subgradient boundedness), we further obtain

$$\begin{aligned}\mathbf{g}_i(t+1)(\mathbf{z}_i(t+1)-\mathbf{v}) \quad &\geq \quad f_i(\mathbf{z}_i(t+1)) - f_i(\bar{\mathbf{x}}(t))\\ &\quad +f_i(\bar{\mathbf{x}}(t)) - f_i(\mathbf{v})\\ &\geq \quad -L_i\|\mathbf{z}_i(t+1)-\bar{\mathbf{x}}(t)\|\\ &\quad +f_i(\bar{\mathbf{x}}(t)) - f_i(\mathbf{v}).\end{aligned}$$

By substituting these estimates back in relation (8), and using $F(\mathbf{x}) = \sum_{i=1}^n f_i(\mathbf{x})$ we obtain

$$\begin{aligned}\sum_{i=1}^n \mathbf{g}_i'(t+1)(\bar{\mathbf{x}}(t)-\mathbf{v}) \quad &\geq \quad F(\bar{\mathbf{x}}(t)) - F(\mathbf{v})\\ &\quad -2\sum_{i=1}^n L_i\|\mathbf{z}_i(t+1)-\bar{\mathbf{x}}(t)\|.\end{aligned} \tag{9}$$

Now, we substitute estimate (9) into relation (7) and obtain for any $\mathbf{v} \in \mathbb{R}^d$ and all $t \geq 0$,

$$\begin{aligned}\|\bar{\mathbf{x}}(t+1)-\mathbf{v}\|^2 \quad &\leq \quad \|\bar{\mathbf{x}}(t)-\mathbf{v}\|^2 - \frac{2\alpha(t+1)}{n}\left(F(\bar{\mathbf{x}}(t))-F(\mathbf{v})\right)\\ &\quad +\frac{4\alpha(t+1)}{n}\sum_{i=1}^n L_i\|\mathbf{z}_i(t+1)-\bar{\mathbf{x}}(t)\|\\ &\quad +\alpha^2(t+1)\frac{(\mathbf{1}'L_F)^2}{n^2}.\end{aligned}$$

□

With all the pieces in place, we are finally ready to prove Theorem 1. The proof idea is to show that the averages $\bar{\mathbf{x}}(t)$, as defined in Lemma 4, converge to some solution $x^* \in X^*$ and then show that $\mathbf{z}_i(t+1) - \bar{\mathbf{x}}(t)$ converges to 0 for all i,

as $t \to \infty$. The last step will be accomplished by invoking Lemma 3 on the perturbed push-sum protocol.

Proof of Theorem 1. We begin by observing that the subgradient-push method may be viewed as an instance of the perturbed push-sum protocol. Indeed, let us adopt the notation $\widetilde{x}_\ell(t), \widetilde{g}_\ell(t)$ from the proof of Lemma 4, and moreover let us define $\widetilde{w}_\ell(t), \widetilde{z}_\ell(t)$ identically. Then, the definition of subgradient-push implies that for all $\ell = 1, \ldots, d$,

$$\begin{aligned}
\widetilde{w}_\ell(t+1) &= A(t)\widetilde{x}_\ell(t), \\
y(t+1) &= A(t)y(t), \\
\widetilde{z}_\ell(t+1) &= \frac{\widetilde{x}_\ell(t)}{y_\ell(t)}, \\
\widetilde{w}_\ell(t+1) &= \widetilde{x}_\ell(t+1) - \alpha(t+1)\widetilde{g}_\ell(t+1).
\end{aligned}$$

Since $\alpha(t) \to 0$, the assumptions of Lemma 3 are satisfied with $\epsilon(t+1) = \alpha(t+1)\widetilde{g}_\ell(t+1)$, from Lemma 3(b) we obtain the conclusion that for all $l = 1, \ldots, d$ and all i,

$$\lim_{t\to\infty} \left| [\widetilde{z}_\ell(t+1)]_i - \frac{\sum_{j=1}^n [\widetilde{x}_\ell(t)]_j}{n} \right| = 0$$

which is equivalent to

$$\lim_{t\to\infty} \|\mathbf{z}_i(t+1) - \bar{\mathbf{x}}(t)\| = 0 \qquad \text{for all } i = 1, \ldots, n. \quad (10)$$

Next, we apply Lemma 3(c). Since the subgradients $\mathbf{g}_i(s)$ are uniformly bounded, and $\{\alpha(t)\}$ is non-increasing and such that $\sum_{t=1}^\infty \alpha^2(t) < \infty$, from $\epsilon(t+1) = \alpha(t+1)\widetilde{g}_\ell(t+1)$ it follows that for all $i = 1, \ldots, n$ and $\ell = 1, \ldots, d$,

$$\begin{aligned}
\sum_{t=1}^\infty \alpha(t)|\epsilon_i(t+1)| &< \sum_{t=1}^\infty \alpha(t)\alpha(t+1)\|\widetilde{g}_\ell(t+1)\|_\infty \\
&\leq \sum_{t=1}^\infty \alpha^2(t)\|\mathbf{g}_i(t+1)\|_\infty < \infty
\end{aligned}$$

In view of the preceding relation and the assumption that the sequence $\{\alpha(t)\}$ is non-increasing, by applying Lemma 3(b) to each coordinate $\ell = 1, \ldots, d$, we obtain that for all $\ell = 1, \ldots, d$ and all $i = 1, \ldots, n$,

$$\sum_{t=0}^\infty \alpha(t+1) \left| [\widetilde{z}_\ell(t+1)]_i - \frac{\sum_{j=1}^n [\widetilde{x}_\ell(t)]_j}{n} \right| < \infty$$

which implies that

$$\sum_{t=0}^\infty \alpha(t+1)\|\mathbf{z}_i(t+1) - \bar{\mathbf{x}}(t)\| < \infty \qquad \text{for all } i = 1, \ldots, n.$$

Next we consider Lemma 4 where we use $\mathbf{v} = z^*$ for some solution $z^* \in Z^*$,

$$\begin{aligned}
\|\bar{\mathbf{x}}(t+1) - z^*\|^2 &\leq \|\bar{\mathbf{x}}(t) - z^*\|^2 - \frac{2\alpha(t+1)}{n}\left(F(\bar{\mathbf{x}}(t)) - F^*\right) \\
&\quad + \frac{4\alpha(t+1)}{n}\sum_{i=1}^n L_i\|\mathbf{z}_i(t+1) - \bar{\mathbf{x}}(t)\| \quad (11) \\
&\quad + \alpha^2(t+1)\frac{(\mathbf{1}'L_F)^2}{n^2}, \quad (12)
\end{aligned}$$

where F^* is the optimal value (i.e., $F^* = F(z^*)$ for any $z^* \in Z^*$). It then follows that

$$\sum_{t=1}^\infty \frac{4\alpha(t+1)}{n}\sum_{i=1}^n L_i\|\mathbf{z}_i(t+1) - \bar{\mathbf{x}}(t)\| < \infty.$$

Also, by assumption we have that $\sum_{t=1}^\infty \alpha(t) = \infty$ and $\sum_{t=1}^\infty \alpha^2(t) < \infty$. It is not hard to see that as a consequence of these facts, $\{\bar{\mathbf{x}}(t)\}$ must converge to some solution $\widehat{z} \in Z^*$. By relation (10) it follows that the sequence $\{\mathbf{z}_i(t)\}$, $i = 1, \ldots, n$, also converges to the solution $\widehat{z}$. □

REFERENCES

[1] F. Benezit, V. Blondel, P. Thiran, J. Tsitsiklis, and M. Vetterli. Weighted gossip: distributed averaging using non-doubly stochastic matrices. In *Proceedings of the 2010 IEEE International Symposium on Information Theory*, Jun. 2010.

[2] J. Chen and A. H. Sayed. Diffusion adaptation strategies for distributed optimization and learning over networks. *IEEE Transactions on Signal Processing*, 60(8):4289–4305, 2012.

[3] A.D. Dominguez-Garcia and C. Hadjicostis. Distributed strategies for average consensus in directed graphs. In *Proceedings of the IEEE Conference on Decision and Control*, Dec 2011.

[4] J.C. Duchi, A. Agarwal, and M.J. Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic Control*, 57(3):592 –606, March 2012.

[5] B. Gharesifard and J. Cortes. Distributed continuous-time convex optimization on weight-balanced digraphs. http://arxiv.org/pdf/1204.0304.pdf, 2012.

[6] B. Johansson. *On distributed optimization in networked systems*. PhD thesis, Royal Institute of Technology (KTH), tRITA-EE 2008:065, 2008.

[7] D. Kempe, A Dobra, and J. Gehrke. Gossip-based computation of aggregate information. In *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science*, pages 482–491, Oct. 2003.

[8] H. Li and Z. Han. Competitive spectrum access in cognitive radio networks: graphical game and learning. In *IEEE Wireless Communications and Networking Conference*, pages 1–6, 2010.

[9] I. Lobel and A. Ozdaglar. Distributed subgradient methods for convex optimization over random networks. *IEEE Transactions on Automatic Control*, 56(6):1291 –1306, June 2011.

[10] I. Lobel, A. Ozdaglar, and D. Feijer. Distributed multi-agent optimization with state-dependent communication. *Mathematical Programming*, 129(2):255–284, 2011.

[11] C. Lopes and A.H. Sayed. Incremental adaptive strategies over distributed networks. *IEEE Transactions on Signal Processing*, 55(8):4046–4077, 2007.

[12] A. Nedić. Asynchronous broadcast-based convex optimizatio over a network. *IEEE Transactions on Automatic Control*, 56(6):1337–1351, 2011.

[13] A. Nedić and A. Ozdaglar. Distributed subgradient methods for multi-agent optimization. *IEEE Transactions on Automatic Control*, 54(1):48 –61, Jan. 2009.

[14] G. Neglia, G. Reina, and S. Alouf. Distributed gradient optimization for epidemic routing: A preliminary evaluation. In *IEEE Wireless Days, 2nd IFIP*, pages 1–6, 2009.

[15] M. Rabbat and R.D. Nowak. Distributed optimization in sensor networks. In *IPSN*, pages 20–27, 2004.

[16] S. S. Ram, A. Nedić, and V. V. Veeravalli. A new class of distributed optimization algorithms: application to regression of distributed data. *Optimization Methods and Software*, 27(1):71–88, 2012.

[17] S.S. Ram, A. Nedić, and V.V. Veeravalli. Distributed Stochastic Subgradient Projection Algorithms for Convex Optimization. *Journal of Optimization Theory and Applications*, 147:516–545, 2010.

[18] S.S. Ram, V.V. Veeravalli, and A. Nedić. Distributed non-autonomous power control through distributed convex optimization. In *IEEE INFOCOM*, pages 3001–3005, 2009.

[19] K. Srivastava and A. Nedić. Distributed asynchronous constrained stochastic optimization. *IEEE J. Sel. Topics. Signal Process.*, 5(4):772–790, 2011.

[20] K.I. Tsianos, S. Lawlor, and M.G. Rabbat. Consensus-based distributed optimization: Practical issues and applications in large-scale machine learning. In *Proceedings of the 50th Allerton Conference on Communication, Control, and Computing*, 2012.

[21] K.I. Tsianos, S. Lawlor, and M.G. Rabbat. Push-sum distributed dual averaging for convex optimization. In *Proceedings of the IEEE Conference on Decision and Control*, 2012.