2021 61st IEEE Conference on Decision and Control (CDC)
December 6-9, 2022. Cancún, Mexico

Subgradient-Push Is of the Optimal Convergence Rate

Yixuan Lin Ji Liu

***Abstract*— The push-sum based subgradient is an important method for distributed convex optimization over unbalanced directed graphs, which is known to converge at a rate of $O(\ln t/\sqrt{t})$. This paper shows that the subgradient-push algorithm actually converges at a rate of $O(1/\sqrt{t})$, which is the same as that of the single-agent subgradient and thus optimal. The proposed tool for analyzing push-sum based algorithms is of independent interest.**

I. Introduction

There are three major information fusion schemes in the vast distributed algorithms literature: consensus via stochastic matrices [1], distributed averaging via doubly stochastic matrices [2], and push-sum via column stochastic matrices [3].[1] Among the three, the push-sum scheme is the only one that is able to not only achieve agreement on the average, but also works for directed graphs, allowing uni-directional communication. Because of this, the push-sum scheme has been widely utilized in various distributed algorithms including distributed optimization [4]–[6], distributed deep learning [7], and distributed reinforcement learning [8]–[10].

The push-sum algorithm was first proposed in [3] and sometimes also called weighted gossip [11], ratio consensus [12], [13], and double linear iteration [14]. Although the analysis of the push-sum algorithm is elegant, the analyses of push-sum based algorithms are often quite complicated, e.g., subgradient-push [5], DEXTRA [15] (a push-sum based variant of the well-known EXTRA algorithm [16]) and Push-DIGing [17]. Actually, all these push-sum based algorithms rely on the pioneering analysis and results in [5].

Distributed optimization originated from the work of [18] and has achieved great success in both theory and practice; see survey papers [19]–[21]. Most existing distributed optimization algorithms require the underlying communication network be described by an undirected graph or a balanced directed graph (a directed graph is balanced if the sum of all in-weights equals the sum of all out-weights at each of its vertices [22]), which allows a distributed manner to construct a doubly stochastic matrix. Such a distributed algorithm usually achieves the same order of convergence rate as its single-agent counterpart, with a difference at a constant coefficient depending on graph connectivity [23].

The push-sum based subgradient algorithm proposed in [5] is the first distributed convex optimization algorithm which works for unbalanced directed graphs without any network-wide information. There are two "gaps" in the analysis in [5]. First, the convergence rate analysis is based on a special convex combination of the history of the states of all agents (see Theorem 2 in [5]), which is "unusual" compared with non-push-sum based distributed optimization algorithms (see e.g. [18]). Second, more importantly, the convergence rate derived in [5] is of order $O(\ln t/\sqrt{t})$, which is slower than that of the single-agent subgradient method, $O(1/\sqrt{t})$ (see Theorem 7 in [23]). With these in mind, this paper aims to close the theoretical gap between the convergence rates of conventional single-agent subgradient and push-sum based subgradient, by analyzing the "standard" convex combination of the history of the states of all agents. We achieve this goal by establishing the explicit "absolute probability sequence" for the push-sum algorithm, which yields a novel analysis tool for push-sum based distributed algorithms over possibly time-varying, unbalanced, directed graphs.

II. Subgradient-Push

Consider a network consisting of n agents, labeled 1 through n for the purpose of presentation. The agents are not aware of such a global labeling, but can differentiate between their neighbors. The neighbor relations among the n agents are characterized by a time-dependent directed graph $\mathbb{G}(t) = (\mathcal{V}, \mathcal{E}(t))$ whose vertices correspond to agents and whose directed edges (or arcs) depict neighbor relations, where $\mathcal{V} = \{1, \ldots, n\}$ is the vertex set and $\mathcal{E}(t) \subset \mathcal{V} \times \mathcal{V}$ is the directed edge set at time t. Specifically, agent j is an in-neighbor of agent i at time t if $(j, i) \in \mathcal{E}(t)$, and similarly, agent k is an out-neighbor of agent i at time t if $(i, k) \in \mathcal{E}(t)$. Each agent can send information to its out-neighbors and receive information from its in-neighbors. Thus, the directions of edges represent the directions of information flow. For convenience, we assume that each agent is always an in- and out-neighbor of itself, which implies that $\mathbb{G}(t)$ has self-arcs at all vertices for all time t. We use $\mathcal{N}_i(t)$ and $\mathcal{N}_i^-(t)$ to denote the in- and out-neighbor set of agent i at time t, respectively, i.e.,

$$\mathcal{N}_i(t) = \{j \in \mathcal{V} : (j, i) \in \mathcal{E}(t)\},$$
$$\mathcal{N}_i^-(t) = \{k \in \mathcal{V} : (i, k) \in \mathcal{E}(t)\}.$$

It is clear that $\mathcal{N}_i(t)$ and $\mathcal{N}_i^-(t)$ are nonempty as they both contain index i. The goal of the n agents is to cooperatively

This material is based upon work supported in part by the National Science Foundation under Grant No. 2230101 and by Stony Brook University's Office of the Vice President for Research through a Seed Grant. The authors wish to thank the anonymous reviewers for their helpful comments.

Y. Lin is with the Department of Applied Mathematics and Statistics at Stony Brook University (yixuan.lin.1@stonybrook.edu). J. Liu is with the Department of Electrical and Computer Engineering at Stony Brook University (ji.liu@stonybrook.edu).

[1]A square nonnegative matrix is called a row stochastic matrix, or simply stochastic matrix, if its row sums all equal one. Similarly, a square nonnegative matrix is called a column stochastic matrix if its column sums all equal one. A square nonnegative matrix is called a doubly stochastic matrix if its row sums and column sums all equal one.

to minimize the cost function

$$f(z) = \frac{1}{n}\sum_{i=1}^{n} f_i(z),$$

where each f_i is a "private" convex (not necessarily differentiable) cost function only known to agent i. It is assumed that the set of optimal solutions to f, denoted by $\mathcal{Z}$, is nonempty.

Since each f_i is not necessarily differentiable, the gradient descent method may not be applicable. Instead, the subgradient method [24] can be applied. For a convex function $h : \mathbb{R}^d \to \mathbb{R}$, a vector $g \in \mathbb{R}^d$ is called a subgradient of h at point x if

$$h(y) \geq h(x) + g^\top (y - x) \ \text{ for all } \ y \in \mathbb{R}^d. \tag{1}$$

Such a vector g always exists and may not be unique. In the case when h is differentiable at point x, the subgradient g is unique and equals $\nabla h(x)$, the gradient of h at x. Thus, the subgradient can be viewed as a generalization of the notion of the gradient. From (1) and the Cauchy-Schwarz inequality,

$$h(y) - h(x) \geq -G\|y - x\|_2, \tag{2}$$

where G is an upper bound for the 2-norm of the subgradients of h at both x and y.

The subgradient method was first proposed in [24] and the first distributed subgradient method was proposed in [18], which is based on average consensus. The subgradient-push algorithm, proposed in [5], is as follows[2]:

$$x_i(t+1) = \sum_{j\in\mathcal{N}_i(t)} w_{ij}(t)\Big[x_j(t) - \alpha(t)g_j(t)\Big], \tag{3}$$

$$y_i(t+1) = \sum_{j\in\mathcal{N}_i(t)} w_{ij}(t)y_j(t), \qquad y_i(0) = 1, \tag{4}$$

where $\alpha(t)$ is the stepsize, $g_j(t)$ is a subgradient of $f_j(z)$ at $x_j(t)/y_j(t)$, and $w_{ij}(t)$, $j \in \mathcal{N}(t)$, are positive weights satisfying the following assumption.

Assumption 1: There exists a constant $\beta > 0$ such that for all $i, j \in \mathcal{V}$ and t, $w_{ij}(t) \geq \beta$ whenever $j \in \mathcal{N}_i(t)$. For all $i \in \mathcal{V}$ and t, $\sum_{j\in\mathcal{N}_i^-(t)} w_{ji}(t) = 1$.

A typical choice of $w_{ij}(t)$ is $1/|\mathcal{N}_j^-(t)|$ for all $j \in \mathcal{N}_i(t)$ which can be computed in a distributed manner and satisfies Assumption 1 with $\beta = 1/n$. Let $W(t)$ be the $n \times n$ matrix whose ijth entry equals $w_{ij}(t)$ if $j \in \mathcal{N}_i(t)$ and zero otherwise; in other words, we set $w_{ij}(t) = 0$ for all $j \notin \mathcal{N}_i(t)$. From Assumption 1, each $W(t)$ is a column stochastic matrix that is compliant with the neighbor graph $\mathbb{G}(t)$. Since each agent i is always assumed to be an in-neighbor of itself, all diagonal entries of $W(t)$ are positive.

To state the convergence result of the subgradient-push algorithm, we need the following assumption and concept.

Assumption 2: The step-size sequence $\{\alpha(t)\}$ is positive, non-increasing, and satisfies $\sum_{t=0}^{\infty}\alpha(t) = \infty$ and $\sum_{t=0}^{\infty}\alpha^2(t) < \infty$.

[2]The algorithm is written in a different but mathematically equivalent form in [5].

Definition 1: A directed graph sequence $\{\mathbb{G}(t)\}$ is uniformly strongly connected if there exists a positive integer L such that for any $t \geq 0$, the union graph $\cup_{k=t}^{t+L-1}\mathbb{G}(k)$ is strongly connected. If such an integer exists, we sometimes say that $\{\mathbb{G}(t)\}$ is uniformly strongly connected by subsequences of length L.

It is not hard to prove that the above definition is equivalent to the two popular joint connectivity definitions in consensus literature, namely "B-connected" [25] and "repeatedly jointly strongly connected" [1].

Define $z_i(t) = x_i(t)/y_i(t)$ and $\bar{z}(t) = \frac{1}{n}\sum_{i=1}^{n} z_i(t)$.

Theorem 1: Suppose that $\{\mathbb{G}_t\}$ is uniformly strongly connected and that $\|g_i(t)\|_2$ is uniformly bounded for all i and t.

1) If the stepsize $\alpha(t)$ is time-varying and satisfies Assumption 2, then with $z^* \in \mathcal{Z}$,

$$\lim_{t\to\infty} f\left(\frac{\sum_{\tau=0}^{t}\alpha(\tau)\bar{z}(\tau)}{\sum_{\tau=0}^{t}\alpha(\tau)}\right) = f(z^*).$$

2) If the stepsize is fixed and $\alpha(t) = 1/\sqrt{T}$ for $T > 0$ steps, i.e., $t \in \{0, 1, \ldots, T-1\}$, then with $z^* \in \mathcal{Z}$,

$$f\left(\frac{\sum_{\tau=0}^{T-1}\bar{z}(\tau)}{T}\right) - f(z^*) \leq O\Big(\frac{1}{\sqrt{T}}\Big).$$

The above theorem establishes the convergence rate of $f((\sum_{\tau=0}^{T-1}\bar{z}(\tau))/T)$, as conventionally did in average-consensus-based subgradient [18], and the rate is of $O(1/\sqrt{t})$, which is the same as that of the conventional single-agent subgradient method [23, Theorem 7]. Thus, the derived convergence rate is optimal.

Theorem 1 is actually a consequence of the following refined result, which further provides finite-time error bounds for the subgradient-push algorithm.

Theorem 2: Suppose that $\{\mathbb{G}_t\}$ is uniformly strongly connected by sub-sequences of length L and that $\|g_i(t)\|_2$ is uniformly bounded above by a positive number G for all i and t.

1) If the stepsize $\alpha(t)$ is time-varying and satisfies Assumption 2, then for all $t \geq 0$,

$$\begin{aligned} &f\left(\frac{\sum_{\tau=0}^{t}\alpha(\tau)\bar{z}(\tau)}{\sum_{\tau=0}^{t}\alpha(\tau)}\right) - f(z^*) \\ &\leq \frac{\|\bar{z}(0) - z^*\|_2^2 + G^2\sum_{\tau=0}^{t}\alpha^2(\tau)}{\sum_{\tau=0}^{t}2\alpha(\tau)} \\ &\quad + \frac{2G\alpha(0)\sum_{i=1}^{n}\|\bar{z}(0) - z_i(0)\|_2}{n\sum_{\tau=0}^{t}\alpha(\tau)} \\ &\quad + \frac{32G}{\eta}\Big\|\sum_{i=1}^{n}x_i(0) + \alpha(0)g_i(0)\Big\|_2\frac{\sum_{\tau=0}^{t-1}\alpha(\tau)\mu^\tau}{\sum_{\tau=0}^{t}\alpha(\tau)} \\ &\quad + \frac{32nG^2}{\eta(1-\mu)}\frac{\sum_{\tau=0}^{t-1}\alpha(\tau)\left(\alpha(0)\mu^{\tau/2} + \alpha(\lceil\frac{\tau}{2}\rceil)\right)}{\sum_{\tau=0}^{t}\alpha(\tau)}. \end{aligned} \tag{5}$$

2) If the stepsize is fixed and $\alpha(t) = 1/\sqrt{T}$ for $T > 0$

steps, i.e., $t \in \{0, 1, \ldots, T-1\}$, then

$$f\Big(\frac{\sum_{\tau=0}^{T-1} \bar{z}(\tau)}{T}\Big) - f(z^*) \leq \frac{2G\sum_{i=1}^n \|\bar{z}(0) - z_i(0)\|_2}{nT} + \frac{\|\bar{z}(0) - z^*\|_2^2 + G^2}{2\sqrt{T}} + \frac{32G}{\eta(1-\mu)T}\Big\|\sum_{i=1}^n x_i(0) + \frac{1}{\sqrt{T}} g_i(0)\Big\|_2 + \frac{32nG^2}{\eta(1-\mu)\sqrt{T}}. \quad (6)$$

Here η and μ are positive constants which satisfy $\eta \geq \frac{1}{n^{nL}}$ and $\mu \leq (1 - \frac{1}{n^{nL}})^{1/L}$, respectively, and $\lceil \cdot \rceil$ denotes the ceiling function.

The above theorem characterizes convergence rates for a network-wide averaged state. The following theorem provides convergence rates for each individual agent.

Theorem 3: Suppose that $\{\mathbb{G}_t\}$ is uniformly strongly connected by sub-sequences of length L and that $\|g_i(t)\|_2$ is uniformly bounded above by a positive number G for all i and t.

1) If the stepsize $\alpha(t)$ is time-varying and satisfies Assumption 2, then for all $t \geq 0$ and $k \in \mathcal{V}$,

$$f\left(\frac{\sum_{\tau=0}^t \alpha(\tau) z_k(\tau)}{\sum_{\tau=0}^t \alpha(\tau)}\right) - f(z^*) \leq \frac{\|\bar{z}(0) - z^*\|_2^2 + G^2 \sum_{\tau=0}^t \alpha^2(\tau)}{\sum_{\tau=0}^t 2\alpha(\tau)} + \frac{G\alpha(0)\sum_{i=1}^n (\|\bar{z}(0) - z_i(0)\|_2 + \|z_k(0) - z_i(0)\|_2)}{n\sum_{\tau=0}^t \alpha(\tau)} + \frac{32G}{\eta}\Big\|\sum_{i=1}^n x_i(0) + \alpha(0) g_i(0)\Big\|_2 \frac{\sum_{\tau=0}^{t-1} \alpha(\tau)\mu^\tau}{\sum_{\tau=0}^t \alpha(\tau)} + \frac{32nG^2}{\eta(1-\mu)} \frac{\sum_{\tau=0}^{t-1} \alpha(\tau)\left(\alpha(0)\mu^{\tau/2} + \alpha(\lceil \frac{\tau}{2} \rceil)\right)}{\sum_{\tau=0}^t \alpha(\tau)}. \quad (7)$$

2) If the stepsize is fixed and $\alpha(t) = 1/\sqrt{T}$ for $T > 0$ steps, i.e., $t \in \{0, 1, \ldots, T-1\}$, then for any $k \in \mathcal{V}$,

$$f\left(\frac{\sum_{\tau=0}^{T-1} z_k(\tau)}{T}\right) - f(z^*) \leq \frac{\|\bar{z}(0) - z^*\|_2^2 + G^2}{2\sqrt{T}} + \frac{32nG^2}{\eta(1-\mu)\sqrt{T}} + \frac{G\sum_{i=1}^n \|\bar{z}(0) - z_i(0)\|_2 + \|z_k(0) - z_i(0)\|_2}{nT} + \frac{32G}{\eta(1-\mu)T}\Big\|\sum_{i=1}^n x_i(0) + \frac{1}{\sqrt{T}} g_i(0)\Big\|_2. \quad (8)$$

Here the positive constants η and $\mu < 1$ are the same as in Theorem 2.

Using the same argument as in the proof of Theorem 1, we have for each agent $k \in \mathcal{V}$, with a time-varying stepsize $\alpha(t)$ satisfying Assumption 2,

$$\lim_{t\to\infty} f\left(\frac{\sum_{\tau=0}^t \alpha(\tau) z_k(\tau)}{\sum_{\tau=0}^t \alpha(\tau)}\right) = f(z^*),$$

and with a fixed stepsize $\alpha(t) = 1/\sqrt{T}$ for $T > 0$ steps,

$$f\Big(\frac{\sum_{\tau=0}^{T-1} z_k(\tau)}{T}\Big) - f(z^*) \leq O\Big(\frac{1}{\sqrt{T}}\Big).$$

III. Analysis

In this section, we provide a novel analysis of the subgradient-push algorithm (3)–(4) and proofs of Theorems 1 and 2. The analysis appeals to the concept of "absolute probability sequence" for push-sum. Thus, we begin with revisiting the well-known push-sum algorithm.

A. Push-Sum

In the push-sum algorithm, each agent i has control over two variables, $x_i(t) \in \mathbb{R}^d$ and $y_i(t) \in \mathbb{R}$, which are updated as follows:

$$x_i(t+1) = \sum_{j\in\mathcal{N}_i(t)} w_{ij}(t) x_j(t), \quad (9)$$

$$y_i(t+1) = \sum_{j\in\mathcal{N}_i(t)} w_{ij}(t) y_j(t), \qquad y_i(0) = 1, \quad (10)$$

where $w_{ij}(t)$, $j \in \mathcal{N}(t)$, are positive weights satisfying Assumption 1.

Let $x(t) \triangleq [x_1(t) \ \cdots \ x_n(t)]^\top \in \mathbb{R}^{n\times d}$ and $y(t)$ be the vector in $\mathbb{R}^n$ whose ith entry is $y_i(t)$. From (9) and (10), $x(t+1) = W(t)x(t)$ and $y(t+1) = W(t)y(t)$. Since $W(t)$ is always column stochastic for all $t \geq 0$, it is easy to show that $\sum_{i=1}^n x_i(t) = \sum_{i=1}^n x_i(0)$ and $\sum_{i=1}^n y_i(t) = \sum_{i=1}^n y_i(0) = n$ for all $t \geq 0$.

Lemma 1: Suppose that $\{\mathbb{G}(t)\}$ is uniformly strongly connected. Then, for any fixed $\tau \geq 0$, $W(t)\cdots W(\tau+1)W(\tau)$ will converge to the set $\{v\mathbf{1}^\top : v \in \mathbb{R}^n, \mathbf{1}^\top v = 1, v > \mathbf{0}\}$ exponentially fast as $t \to \infty$.[3]

The lemma is essentially the same as Corollary 2 (a) in [5]. Suppose $\{\mathbb{G}_t\}$ is uniformly strongly connected by sub-sequences of length L, Lemma 1 implies that there exist constants $c > 0$ and $\mu \in [0, 1)$ and a sequence of stochastic vectors[4] $\{v(t)\}$ such that for all $i, j \in \mathcal{V}$ and $t \geq \tau \geq 0$,

$$\big|\big[W(t)\cdots W(\tau+1)W(\tau)\big]_{ij} - v_i(t)\big| \leq c\mu^{t-\tau}, \quad (11)$$

where $[\cdot]_{ij}$ denotes the ijth entry of a matrix. In [5], it has been shown that $c = 4$ and $\mu = (1 - \frac{1}{n^{nL}})^{1/L}$.

To proceed, we define a time-dependent $n \times n$ matrix $S(t)$ whose ijth entry is

$$s_{ij}(t) = \frac{w_{ij}(t)y_j(t)}{y_i(t+1)} = \frac{w_{ij}(t)y_j(t)}{\sum_{k=1}^n w_{ik}(t)y_k(t)}. \quad (12)$$

[3]We use $\mathbf{0}$ and $\mathbf{1}$ to denote the vectors whose entries all equal to 0 or 1, respectively, where the dimensions of the vectors are to be understood from the context. We use $v > \mathbf{0}$ to denote a positive vector, i.e., each entry of v is positive.

[4]A vector is called a stochastic vector if its entries are all nonnegative and sum to one.

It is worth emphasizing that $S(t)$ is independent of $x(t)$. The following lemma guarantees that $S(t)$ is well defined.

Lemma 2: Suppose that $\{\mathbb{G}(t)\}$ is uniformly strongly connected, then there exists a constant $\eta > 0$ such that $n \geq y_i(t) \geq \eta$ for all i and t.

The lemma is essentially the same as Corollary 2 (b) in [5], which further proves that if $\{\mathbb{G}_t\}$ is uniformly strongly connected by sub-sequences of length L, then $\eta \geq \frac{1}{n^{nL}}$.

Define $z_i(t) = x_i(t)/y_i(t)$ for each $i \in \mathcal{V}$. Then,

$$\begin{aligned} z_i(t+1) &= \frac{x_i(t+1)}{y_i(t+1)} = \frac{\sum_{j=1}^n w_{ij}(t)x_j(t)}{\sum_{j=1}^n w_{ij}(t)y_j(t)} \\ &= \sum_{j=1}^n \frac{w_{ij}(t)x_j(t)}{\sum_{k=1}^n w_{ik}(t)y_k(t)} = \sum_{j=1}^n \left[\frac{w_{ij}(t)y_j(t)}{\sum_{k=1}^n w_{ik}(t)y_k(t)}\right] z_j(t) \\ &= \sum_{j=1}^n s_{ij}(t)z_j(t), \end{aligned} \tag{13}$$

which implies that $z(t+1) = S(t)z(t)$ where $z(t) \triangleq [z_1(t) \ \cdots \ z_n(t)]^\top \in \mathbb{R}^{n\times d}$. Actually $S(t)$ is always a stochastic matrix, as we will show shortly.

Similar to the discrete-time state transition matrix, let $\Phi_W(t,\tau) = W(t-1)\cdots W(\tau)$ with $t > \tau$, and similarly, let $\Phi_S(t,\tau) = S(t-1)\cdots S(\tau)$ with $t > \tau$.

Lemma 3: For $i, j \in \mathcal{V}$ and $t > \tau \geq 0$, there holds $[\Phi_S(t,\tau)]_{ij}y_i(t) = [\Phi_W(t,\tau)]_{ij}y_j(\tau)$.

Proof of Lemma 3: The claim will be proved by induction on t. For the basis step, suppose that $t = \tau + 1$. Then, from (12), $[\Phi_S(\tau+1,\tau)]_{ij} = s_{ij}(\tau) = \frac{y_j(\tau)w_{ij}(\tau)}{y_i(\tau+1)} = \frac{y_j(\tau)}{y_i(\tau+1)}[\Phi_W(\tau+1,\tau)]_{ij}$. Thus, in this case the claim is true. For the inductive step, suppose that the claim holds for $t = h > \tau$, where h is a positive integer, and that $t = h+1$. Then,

$$\begin{aligned} [\Phi_S(h+1,\tau)]_{ij} &= \sum_{k=1}^n s_{ik}(h)\cdot[\Phi_S(h,\tau)]_{kj} \\ &= \sum_{k=1}^n \frac{w_{ik}(h)y_k(h)}{y_i(h+1)}\cdot\frac{y_j(\tau)}{y_k(h)}[\Phi_W(h,\tau)]_{kj} \\ &= \frac{y_j(\tau)}{y_i(h+1)}\sum_{k=1}^n w_{ik}(h)\cdot[\Phi_W(h,\tau)]_{kj} \\ &= \frac{y_j(\tau)}{y_i(h+1)}[\Phi_W(h+1,\tau)]_{ij}, \end{aligned}$$

which establishes the claim by induction. ■

More can be said.

Lemma 4: Suppose that $\{\mathbb{G}(t)\}$ is uniformly strongly connected. Then, for any fixed $\tau \geq 0$, $S(t)\cdots S(\tau+1)S(\tau)$ will converge to $\frac{1}{n}\mathbf{1}y^\top(\tau)$.

Proof of Lemma 4: From Lemma 1, for any given $\tau \geq 0$, there holds $\lim_{t\to\infty}[\Phi_W(t,\tau)] = v(\tau,\infty)\mathbf{1}^\top$, with the understanding that $v(\tau,\infty)$ is not necessarily a constant vector. From Lemma 3 and the fact that $y(t) = \Phi_W(t,\tau)y(\tau)$ for all $t > \tau$, for any $i, j \in \mathcal{V}$ we have

$$\begin{aligned} &\lim_{t\to\infty}[\Phi_S(t,\tau)]_{ij} \\ &= \lim_{t\to\infty}\frac{y_j(\tau)}{y_i(t)}[\Phi_W(t,\tau)]_{ij} = \lim_{t\to\infty}\frac{y_j(\tau)[\Phi_W(t,\tau)]_{ij}}{\sum_{k=1}^n[\Phi_W(t,\tau)]_{ik}y_k(\tau)} \\ &= \frac{y_j(\tau)\lim_{t\to\infty}[\Phi_W(t,\tau)]_{ij}}{\lim_{t\to\infty}\sum_{k=1}^n[\Phi_W(t,\tau)]_{ik}y_k(\tau)} \\ &= \frac{y_j(\tau)v_i(\tau,\infty)}{\sum_{k=1}^n v_i(\tau,\infty)y_k(\tau)} \overset{(a)}{=} \frac{y_j(\tau)}{\sum_{k=1}^n y_k(\tau)} \overset{(b)}{=} \frac{y_j(\tau)}{n}, \end{aligned}$$

where in (a) we used the fact that $v(\tau,\infty) > \mathbf{0}$ by Lemma 1 and in (b) we used the fact that $\sum_{i=1}^n y_i(t) = n$ for all $t \geq 0$. ■

Proposition 1: Suppose that $\{\mathbb{G}(t)\}$ is uniformly strongly connected. Then, for any fixed $\tau \geq 0$, $S(t)\cdots S(\tau+1)S(\tau)$ will converge to $\frac{1}{n}\mathbf{1}y^\top(\tau)$ exponentially fast as $t \to \infty$.

Proof of Proposition 1: The proof can be found in [26]. ■

The proposition immediately implies the following results.

Corollary 1: Suppose that $\{\mathbb{G}(t)\}$ is uniformly strongly connected. Then, $S(t)\cdots S(1)S(0)$ will converge to $\frac{1}{n}\mathbf{1}\mathbf{1}^\top$ exponentially fast as $t \to \infty$.

Proof of Corollary 1: The corollary is a special case of Proposition 1 by setting $\tau = 0$. ■

Corollary 2: If $\{\mathbb{G}(t)\}$ is uniformly strongly connected, then $x_i(t)/y_i(t)$ for all $i \in \mathcal{V}$ converges to $\frac{1}{n}\sum_{i=1}^n x_i(0)$ exponentially fast.

Proof of Corollary 2: The proof can be found in [26]. ■

Although the proof of Corollary 2 looks more complicated than the conventional convergence proof of the push-sum algorithm (e.g., [3], [12], [14]), it yields the following novel and key property of push-sum.

To proceed, we rewrite the push-sum algorithm in a different form which directly characterizes the dynamics of $z_i(t) = x_i(t)/y_i(t)$. From (12) and (13), $z_i(t+1) = \sum_{j=1}^n s_{ij}(t)z_j(t)$ and $s_{ij}(t)$ satisfies the following assumption.

Assumption 3: There exists a constant $\gamma > 0$ such that for all $i, j \in \mathcal{V}$ and t, $s_{ii}(t) \geq \gamma$ and $s_{ij}(t) \geq \gamma$ whenever $s_{ij}(t) > 0$. For all $i \in \mathcal{V}$ and t, $\sum_{j=1}^n s_{ij}(t) = 1$.

Lemma 5: Suppose that Assumption 1 holds. Then, $s_{ij}(t)$ satisfies Assumption 3 for each $t \geq 0$.

Proof of Lemma 5: From Assumption 1, each $W(t)$ is a column stochastic matrix whose diagonal entries are all positive and $w_{ij}(t) \geq \beta$ whenever $w_{ij}(t) > 0$. From (12), $s_{ij}(t) > 0$ only if $w_{ij}(t) > 0$. From Lemma 2, when $w_{ij}(t) > 0$,

$$s_{ij}(t) = \frac{w_{ij}(t)y_j(t)}{y_i(t+1)} \geq \frac{\beta\eta}{n}.$$

The above inequality and Assumption 1 imply that $s_{ij}(t)$ satisfies the first condition of Assumption 3 with $\gamma = \beta\eta/n$.

For the second condition of Assumption 3, it is clear that

$$\sum_{j=1}^{n} s_{ij}(t) = \sum_{j=1}^{n} \frac{w_{ij}(t)y_j(t)}{\sum_{k=1}^{n} w_{ik}(t)y_k(t)} = 1$$

for all $i \in \mathcal{V}$ and t, which completes the proof. ■

From Lemma 5, each $S(t)$ is a row stochastic matrix whose diagonal entries are all positive and whose nonzero entries are all uniformly bounded below by some positive number. More can be said. The following lemma shows that each $S(t)$ is compliant with the neighbor graph $\mathbb{G}(t)$.

Lemma 6: The graph of $S(t)$ is the same as the graph of $W(t)$ for all t.[5]

Proof of Lemma 6: From (12) and Lemma 2, it is easy to see that $s_{ij}(t) > 0$ if and only if $w_{ij}(t)$, which proves the lemma. ■

From (13), $z(t+1) = S(t)z(t)$. The above lemmas imply that the dynamics of $z(t)$ is a nonlinear consensus process as $S(t)$ is dependent on $z(t)$. Such a transition in analysis from $x(t)$ dynamics to $z(t)$ dynamics has been used in [27]. To analyze such a process, we appeal to the following concept, which has been applied to analyze consensus processes [28], [29] and consensus-based distributed optimization [30]. To our knowledge, the concept has never been used to analyze the push-sum algorithm and its applications.

Definition 2: Let $\{S(t)\}$ be a sequence of stochastic matrices. A sequence of stochastic vectors $\{\pi(t)\}$ is an absolute probability sequence for $\{S(t)\}$ if $\pi^\top(t) = \pi^\top(t+1)S(t)$ for all $t \geq 0$.

This definition was first introduced by Kolmogorov [31]. It was shown by Blackwell [32] that every sequence of stochastic matrices has an absolute probability sequence. In general, a sequence of stochastic matrices may have more than one absolute probability sequence; when the sequence of stochastic matrices is "ergodic",[6] it has a unique absolute probability sequence [29, Lemma 1]. It is easy to see that when $S(t)$ is a fixed irreducible stochastic matrix S, $\pi(t)$ is simply the normalized left eigenvector of S for eigenvalue one, and when $\{S(t)\}$ is an ergodic sequence of doubly stochastic matrices, $\pi(t) = (1/n)\mathbf{1}$. More can be said.

Lemma 7: (Theorem 4.8 in [28]) Let $\{S(t)\}$ be a sequence of stochastic matrices satisfying Assumption 3. If the graph sequence of $\{\mathbb{G}(t)\}$ is uniformly strongly connected, then there exists a unique absolute probability sequence $\{\pi(t)\}$ for the matrix sequence $\{S(t)\}$ and a constant $\pi_{\min} \in (0,1)$ such that $\pi_i(t) \geq \pi_{\min}$ for all i and t.

A particular important property of the absolute probability sequence for $\{S(t)\}$ is as follows.

Proposition 2: Suppose that $\{\mathbb{G}(t)\}$ is uniformly strongly connected. Then, the sequence of stochastic matrices $\{S(t)\}$ has a unique absolute probability sequence $\{\pi(t)\}$ with $\pi_i(t) = \frac{y_i(t)}{n}$ for all $i \in \mathcal{V}$ and $t \geq 0$.

The proposition is a consequence of Lemma 1 in [29]. We provide two alternative proofs.

Proof of Proposition 2: First, Lemma 4 shows that $\{S(t)\}$ is ergodic, so it must have a unique absolute probability sequence $\{\pi(t)\}$. From Definition 2 and Lemma 4, for any $\tau \geq 0$,

$$\begin{aligned}\pi^\top(\tau) &= \pi^\top(\tau+1)S(\tau) = \pi^\top(\tau+2)S(\tau+1)S(\tau)\\ &= \lim_{t\to\infty} \pi^\top(t+1)S(t)\cdots S(\tau+1)S(\tau)\\ &= \lim_{t\to\infty} \frac{1}{n}\pi^\top(t+1)\mathbf{1}y^\top(\tau) = \frac{1}{n}y^\top(\tau),\end{aligned}$$

which proves the statement.

Alternatively, we can also prove the proposition by showing that the sequence $\{\pi(t)\}$ with $\pi_i(t) = \frac{y_i(t)}{n}$ satisfies $\pi^\top(t) = \pi^\top(t+1)S(t)$. To see this, from (12) and Assumption 1, for $j \in \mathcal{V}$

$$\begin{aligned}[\pi^\top(t+1)S(t)]_j &= \sum_{i=1}^{n} \frac{y_i(t+1)}{n} s_{ij}(t)\\ = \sum_{i=1}^{n} \frac{y_i(t+1)}{n} \frac{w_{ij}(t)y_j(t)}{y_i(t+1)} &= \sum_{i=1}^{n} \frac{w_{ij}(t)y_j(t)}{n}\\ = \frac{y_j(t)}{n} &= \pi_j^\top(t).\end{aligned}$$

This completes the proof. ■

Next we will appeal to this property to construct a novel time-varying Lyapunov function for distributed convex optimization which yields an improved convergence rate of the subgradient-push algorithm.

Remark 1: Since the stochastic matrix sequence $S(t)$ defined by (12) is purely based on the $y_i(t)$ variables and is thus independent of the $x_i(t)$ variables of the push-sum algorithm, so its absolute probability sequence. Considering the fact that the push-sum and subgradient-push algorithms share the same $y_i(t)$ dynamics which is independent of their $x_i(t)$ dynamics, all the results of $\{S(t)\}$ and its absolute probability sequence derived in this subsection also apply to the subgradient-push algorithm. □

B. Subgradient-Push

We first rewrite the subgradient-push algorithm as follows. From (3)–(4), we have

$$\begin{aligned}z_i(t+1) = \frac{x_i(t+1)}{y_i(t+1)} &= \frac{\sum_{j=1}^{n} w_{ij}(t)[x_j(t) - \alpha(t)g_j(t)]}{\sum_{j=1}^{n} w_{ij}(t)y_j(t)}\\ &= \sum_{j=1}^{n} \frac{w_{ij}(t)[x_j(t) - \alpha(t)g_j(t)]}{\sum_{k=1}^{n} w_{ik}(t)y_k(t)}\\ &= \sum_{j=1}^{n} \left[\frac{w_{ij}(t)y_j(t)}{\sum_{k=1}^{n} w_{ik}(t)y_k(t)}\right]\left[z_j(t) - \alpha(t)\frac{g_j(t)}{y_j(t)}\right]\\ &= \sum_{j=1}^{n} s_{ij}(t)\left[z_j(t) - \alpha(t)\frac{g_j(t)}{y_j(t)}\right],\end{aligned}$$

[5]The graph of an $n \times n$ matrix is a direct graph with n vertices and an arc from vertex i to vertex j whenever the jith entry of the matrix is nonzero.

[6]A sequence of stochastic matrices $\{S(t)\}$ is called ergodic if $\lim_{t\to\infty} S(t)\cdots S(\tau+1)S(\tau) = \mathbf{1}v^\top(\tau)$ for all τ, where each $v(\tau)$ is a stochastic vector.

where $s_{ij}(t)$ is defined in (12). In addition,

$$\begin{aligned}\bar{z}(t+1) &= \frac{1}{n}\sum_{i=1}^{n} z_i(t+1)\\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} s_{ij}(t)\Big[z_j(t) - \alpha(t)\frac{g_j(t)}{y_j(t)}\Big].\end{aligned}$$

Define a time-varying Lyapunov function

$$\langle z(t)\rangle = \pi^\top(t) z(t).$$

Then, from Definition 2, we have

$$\begin{aligned}&\langle z(t+1)\rangle = \sum_{i=1}^{n}\pi_i(t+1) z_i(t+1)\\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\pi_i(t+1)s_{ij}(t)\Big[z_j(t) - \alpha(t)\frac{g_j(t)}{y_j(t)}\Big]\\ &= \sum_{j=1}^{n}\pi_j(t)\Big[z_j(t) - \alpha(t)\frac{g_j(t)}{y_j(t)}\Big] = \langle z(t)\rangle - \frac{\alpha(t)}{n}\sum_{i=1}^{n} g_i(t),\end{aligned} \tag{14}$$

where we use the Proposition 2 in the last equality.

To prove Theorem 1, we need the following lemma.

Lemma 8: Suppose that $\{\mathbb{G}_t\}$ is uniformly strongly connected by sub-sequences of length L and that $\|g_i(t)\|_2$ is uniformly bounded above by a positive number G for all i and t. Then, for all $t \geq 0$ and $i \in \mathcal{V}$,

$$\begin{aligned}&\Big\|z_i(t+1) - \frac{1}{n}\sum_{j=1}^{n}(x_j(t) - \alpha(t)g_j(t))\Big\|_2\\ &\leq \frac{8}{\eta}\mu^t\Big\|\sum_{i=1}^{n} x_i(0) + \alpha(0)g_i(0)\Big\|_2 + \frac{8nG}{\eta}\sum_{s=1}^{t}\mu^{t-s}\alpha(s).\end{aligned}$$

Suppose, in addition, that Assumption 2 holds. Then, for all $t \geq 0$ and $i \in \mathcal{V}$,

$$\begin{aligned}&\Big\|z_i(t+1) - \frac{1}{n}\sum_{j=1}^{n}(x_j(t) - \alpha(t)g_j(t))\Big\|_2\\ &\leq \frac{8}{\eta}\mu^t\Big\|\sum_{i=1}^{n} x_i(0) + \alpha(0)g_i(0)\Big\|_2\\ &\quad + \frac{8nG}{\eta(1-\mu)}\Big(\alpha(0)\mu^{t/2} + \alpha(\lceil t/2\rceil)\Big).\end{aligned}$$

Here $\eta > 0$ and $\mu \in (0,1)$ are defined in Lemma 2 and (11), respectively.

Proof of Lemma 8: The proof can be found in [26]. ■

We are now in a position to prove Theorem 2.

Proof of Theorem 2: Note that for all $t \geq 0$ and $i \in \mathcal{V}$,

$$\begin{aligned}&\big\|\langle z(t+1)\rangle - z_i(t+1)\big\|_2 + \big\|\bar{z}(t+1) - z_i(t+1)\big\|_2\\ &\leq \Big\|\langle z(t+1)\rangle - \frac{1}{n}\sum_{k=1}^{n}(x_k(t) - \alpha(t)g_k(t))\Big\|_2\\ &\quad + \Big\|\bar{z}(t+1) - \frac{1}{n}\sum_{k=1}^{n}(x_k(t) - \alpha(t)g_k(t))\Big\|_2\\ &\quad + 2\Big\|z_i(t+1) - \frac{1}{n}\sum_{k=1}^{n}(x_k(t) - \alpha(t)g_k(t))\Big\|_2\\ &= \Big\|\sum_{j=1}^{n}\pi_j\big(z_j(t+1) - \frac{1}{n}\sum_{k=1}^{n}(x_k(t) - \alpha(t)g_k(t))\big)\Big\|_2\\ &\quad + \Big\|\sum_{j=1}^{n}\frac{1}{n}\big(z_j(t+1) - \frac{1}{n}\sum_{k=1}^{n}(x_k(t) - \alpha(t)g_k(t))\big)\Big\|_2\\ &\quad + 2\Big\|z_i(t+1) - \frac{1}{n}\sum_{k=1}^{n}(x_k(t) - \alpha(t)g_k(t))\Big\|_2\\ &\leq \sum_{j=1}^{n}\pi_j(t)\Big\|z_j(t+1) - \frac{1}{n}\sum_{k=1}^{n}(x_k(t) - \alpha(t)g_k(t))\Big\|_2\\ &\quad + \sum_{j=1}^{n}\frac{1}{n}\Big\|z_j(t+1) - \frac{1}{n}\sum_{k=1}^{n}(x_k(t) - \alpha(t)g_k(t))\Big\|_2\\ &\quad + 2\Big\|z_i(t+1) - \frac{1}{n}\sum_{k=1}^{n}(x_k(t) - \alpha(t)g_k(t))\Big\|_2\\ &\leq \frac{32}{\eta}\mu^t\Big\|\sum_{i=1}^{n} x_i(0) + \alpha(0)g_i(0)\Big\|_2 + \frac{32nG}{\eta}\sum_{s=0}^{t}\mu^{t-s}\alpha(s),\end{aligned} \tag{15}$$

where we used Lemma 8 in the last inequality. Similarly, in the case when $\{\alpha(t)\}$ satisfies Assumption 2,

$$\begin{aligned}&\big\|\langle z(t+1)\rangle - z_i(t+1)\big\|_2 + \big\|\bar{z}(t+1) - z_i(t+1)\big\|_2\\ &\leq \frac{32}{\eta}\mu^t\Big\|\sum_{i=1}^{n} x_i(0) + \alpha(0)g_i(0)\Big\|_2\\ &\quad + \frac{32nG}{\eta(1-\mu)}\Big(\alpha(0)\mu^{t/2} + \alpha(\lceil t/2\rceil)\Big).\end{aligned} \tag{16}$$

From (14),

$$\begin{aligned}&\big\|\langle z(t+1)\rangle - z^*\big\|_2^2\\ &= \Big\|\langle z(t)\rangle - z^* - \frac{\alpha(t)}{n}\sum_{i=1}^{n} g_i(t)\Big\|_2^2\\ &\leq \big\|\langle z(t)\rangle - z^*\big\|_2^2 + \Big\|\frac{\alpha(t)}{n}\sum_{i=1}^{n} g_i(t)\Big\|_2^2\\ &\quad - 2\big(\langle z(t)\rangle - z^*\big)^\top\Big(\frac{\alpha(t)}{n}\sum_{i=1}^{n} g_i(t)\Big)\\ &\leq \big\|\langle z(t)\rangle - z^*\big\|_2^2 + \alpha^2(t)G^2\\ &\quad - 2\big(\langle z(t)\rangle - z^*\big)^\top\Big(\frac{\alpha(t)}{n}\sum_{i=1}^{n} g_i(t)\Big).\end{aligned} \tag{17}$$

Moreover,

$$\begin{aligned}
&\big(\langle z(t)\rangle - z^*\big)^\top g_i(t)\\
&= \big(\langle z(t)\rangle - z_i(t)\big)^\top g_i(t) + \big(z_i(t) - z^*\big)^\top g_i(t)\\
&\geq f_i(z_i(t)) - f_i(z^*) - G\big\|\langle z(t)\rangle - z_i(t)\big\|_2 \qquad (18)\\
&\geq f_i(\bar z(t)) - f_i(z^*) - G\big\|\langle z(t)\rangle - z_i(t)\big\|_2\\
&\quad - G\big\|\bar z(t) - z_i(t)\big\|_2, \qquad (19)
\end{aligned}$$

where we used (1) and (2) in deriving (18), and made use of (2) to get (19). Combining (17) and (19),

$$\begin{aligned}
&\big\|\langle z(t+1)\rangle - z^*\big\|_2^2\\
&\leq \big\|\langle z(t)\rangle - z^*\big\|_2^2 + \alpha^2(t)G^2 - 2\alpha(t)\big(f(\bar z(t)) - f(z^*)\big)\\
&\quad + \frac{2G\alpha(t)}{n}\sum_{i=1}^n \big(\big\|\langle z(t)\rangle - z_i(t)\big\|_2 + \big\|\bar z(t) - z_i(t)\big\|_2\big),
\end{aligned}$$

which implies that

$$\begin{aligned}
&2\alpha(t)\big(f(\bar z(t)) - f(z^*)\big)\\
&\leq \big\|\langle z(t)\rangle - z^*\big\|_2^2 + \alpha^2(t)G^2 - \big\|\langle z(t+1)\rangle - z^*\big\|_2^2\\
&\quad + \frac{2G\alpha(t)}{n}\sum_{i=1}^n \big(\big\|\langle z(t)\rangle - z_i(t)\big\|_2 + \big\|\bar z(t) - z_i(t)\big\|_2\big).
\end{aligned}$$

Summing this relation over time, it follows that

$$\begin{aligned}
&\sum_{\tau=0}^t 2\alpha(\tau)\big(f(\bar z(\tau)) - f(z^*)\big)\\
&\leq \big\|\langle z(0)\rangle - z^*\big\|_2^2 - \big\|\langle z(t+1)\rangle - z^*\big\|_2^2 + \sum_{\tau=0}^t \alpha^2(\tau)G^2\\
&+ \sum_{\tau=0}^t \frac{2G\alpha(\tau)}{n}\sum_{i=1}^n \big(\big\|\langle z(\tau)\rangle - z_i(\tau)\big\|_2 + \big\|\bar z(\tau) - z_i(\tau)\big\|_2\big).
\end{aligned}$$

Note that

$$\begin{aligned}
&f\Big(\frac{\sum_{\tau=0}^t \alpha(\tau)\bar z(\tau)}{\sum_{\tau=0}^t \alpha(\tau)}\Big) - f(z^*)\\
&\leq \frac{\sum_{\tau=0}^t 2\alpha(\tau)(f(\bar z(\tau)) - f(z^*))}{\sum_{\tau=0}^t 2\alpha(\tau)}.
\end{aligned}$$

It follows that

$$\begin{aligned}
&f\Big(\frac{\sum_{\tau=0}^t \alpha(\tau)\bar z(\tau)}{\sum_{\tau=0}^t \alpha(\tau)}\Big) - f(z^*)\\
&\leq \frac{\|\langle z(0)\rangle - z^*\|_2^2 - \|\langle z(t+1)\rangle - z^*\|_2^2 + \sum_{\tau=0}^t \alpha^2(\tau)G^2}{\sum_{\tau=0}^t 2\alpha(\tau)}\\
&+ \frac{\sum_{\tau=0}^t \frac{2G\alpha(\tau)}{n}\sum_{i=1}^n(\|\langle z(\tau)\rangle - z_i(\tau)\|_2 + \|\bar z(\tau) - z_i(\tau)\|_2)}{\sum_{\tau=0}^t 2\alpha(\tau)}\\
&\leq \frac{\sum_{\tau=0}^t G\alpha(\tau)\sum_{i=1}^n(\|\langle z(\tau)\rangle - z_i(\tau)\|_2 + \|\bar z(\tau) - z_i(\tau)\|_2)}{n\sum_{\tau=0}^t \alpha(\tau)}\\
&+ \frac{\|\langle z(0)\rangle - z^*\|_2^2 + \sum_{\tau=0}^t \alpha^2(\tau)G^2}{\sum_{\tau=0}^t 2\alpha(\tau)}. \qquad (20)
\end{aligned}$$

We next consider the time-varying and fixed stepsizes separately.

1) If the stepsize $\alpha(t)$ is time-varying and satisfies Assumption 2, then combining (16) and (20),

$$\begin{aligned}
&f\Big(\frac{\sum_{\tau=0}^t \alpha(\tau)\bar z(\tau)}{\sum_{\tau=0}^t \alpha(\tau)}\Big) - f(z^*)\\
&\leq \frac{\|\langle z(0)\rangle - z^*\|_2^2 + G^2\sum_{\tau=0}^t \alpha^2(\tau)}{\sum_{\tau=0}^t 2\alpha(\tau)}\\
&\quad + \frac{G\alpha(0)\sum_{i=1}^n(\|\langle z(0)\rangle - z_i(0)\|_2 + \|\bar z(0) - z_i(0)\|_2)}{n\sum_{\tau=0}^t \alpha(\tau)}\\
&\quad + \frac{32G}{\eta}\Big\|\sum_{i=1}^n x_i(0) + \alpha(0)g_i(0)\Big\|_2\frac{\sum_{\tau=0}^{t-1}\alpha(\tau)\mu^\tau}{\sum_{\tau=0}^t \alpha(\tau)}\\
&\quad + \frac{32nG^2}{\eta(1-\mu)}\frac{\sum_{\tau=0}^{t-1}\alpha(\tau)(\alpha(0)\mu^{\tau/2} + \alpha(\lceil\frac{\tau}{2}\rceil))}{\sum_{\tau=0}^t \alpha(\tau)}.
\end{aligned}$$

From Proposition 2 and $y_i(0) = 1$, $\pi_i(0) = \frac{1}{n}$ for all $i \in \mathcal{V}$, which implies that $\langle z(0)\rangle = \frac{1}{n}\sum_i^n z_i(0) = \bar z(0)$. We thus have derived (5).

2) If the stepsize is fixed and $\alpha(t) = 1/\sqrt{T}$ for all $t \geq 0$, then from (20),

$$\begin{aligned}
&f\Big(\frac{\sum_{\tau=0}^{T-1}\bar z(\tau)}{T}\Big) - f(z^*)\\
&\leq \frac{G\sum_{\tau=0}^{T-1}\sum_{i=1}^n(\|\langle z(\tau)\rangle - z_i(\tau)\|_2 + \|\bar z(\tau) - z_i(\tau)\|_2)}{nT}\\
&\quad + \frac{\|\langle z(0)\rangle - z^*\|_2^2 + G^2}{2\sqrt{T}}\\
&\overset{(a)}{\leq} \frac{G\sum_{i=1}^n(\|\langle z(0)\rangle - z_i(0)\|_2 + \|\bar z(0) - z_i(0)\|_2)}{nT}\\
&\quad + \frac{\|\langle z(0)\rangle - z^*\|_2^2 + G^2}{2\sqrt{T}} + \frac{32nG^2}{T\eta}\sum_{\tau=0}^{T-2}\sum_{s=0}^{\tau}\mu^{\tau-s}\frac{1}{\sqrt{T}}\\
&\quad + \frac{32G}{T\eta}\Big\|\sum_{i=1}^n x_i(0) + \frac{1}{\sqrt{T}}g_i(0)\Big\|_2\sum_{\tau=0}^{T-2}\mu^\tau\\
&\leq \frac{G\sum_{i=1}^n(\|\langle z(0)\rangle - z_i(0)\|_2 + \|\bar z(0) - z_i(0)\|_2)}{nT}\\
&\quad + \frac{\|\langle z(0)\rangle - z^*\|_2^2 + G^2}{2\sqrt{T}} + \frac{32nG^2}{\sqrt{T}\eta(1-\mu)}\\
&\quad + \frac{32G}{T\eta(1-\mu)}\Big\|\sum_{i=1}^n x_i(0) + \frac{1}{\sqrt{T}}g_i(0)\Big\|_2,
\end{aligned}$$

where we used (15) in (a). Since $\langle z(0)\rangle = \frac{1}{n}\sum_i^n z_i(0) = \bar z(0)$, we have derived (6). ∎

We next prove Theorem 1.

Proof of Theorem 1: We consider the time-varying and fixed stepsizes separately.

1) If the stepsize $\alpha(t)$ is time-varying and satisfies Assumption 2, then

$$\begin{aligned}
&\lim_{t\to\infty}\frac{\|\langle z(0)\rangle - z^*\|_2^2 + \sum_{\tau=0}^t \alpha^2(\tau)G^2}{\sum_{\tau=0}^t 2\alpha(\tau)} = 0,\\
&\lim_{t\to\infty}\frac{\sum_{i=1}^n(\|\langle z(0)\rangle - z_i(0)\|_2 + \|\bar z(0) - z_i(0)\|_2)}{\sum_{\tau=0}^t \alpha(\tau)} = 0.
\end{aligned}$$

Note that $\sum_{\tau=0}^{t-1}\alpha(\tau)\mu^{\tau}\leq\frac{\alpha(0)}{1-\mu}$ and

$$\sum_{\tau=0}^{t-1}\alpha(\tau)\big(\alpha(0)\mu^{\tau/2}+\alpha(\lceil\frac{\tau}{2}\rceil)\big)\leq\frac{\alpha^2(0)}{1-\mu^{1/2}}+\sum_{\tau=0}^{t-1}\alpha^2(\lceil\frac{\tau}{2}\rceil).$$

It follows that

$$\lim_{t\to\infty}\frac{\sum_{\tau=0}^{t-1}\alpha(\tau)\mu^{\tau}}{\sum_{\tau=0}^{t}\alpha(\tau)}=0,$$
$$\lim_{t\to\infty}\frac{\sum_{\tau=0}^{t-1}\alpha(\tau)(\alpha(0)\mu^{\tau/2}+\alpha(\lceil\frac{\tau}{2}\rceil))}{\sum_{\tau=0}^{t}\alpha(\tau)}=0.$$

From (5),

$$\lim_{t\to\infty}f\bigg(\frac{\sum_{\tau=0}^{t}\alpha(\tau)\bar{z}(\tau)}{\sum_{\tau=0}^{t}\alpha(\tau)}\bigg)-f(z^*)=0.$$

2) If the stepsize is fixed and $\alpha(t)=1/\sqrt{T}$ for all $t\geq 0$, then from (6),

$$f\bigg(\frac{\sum_{\tau=0}^{T-1}\bar{z}(\tau)}{T}\bigg)-f(z^*)\leq O\Big(\frac{1}{\sqrt{T}}\Big).$$

This completes the proof. ■

We finally prove Theorem 3.

Proof of Theorem 3: The proof can be found in [26]. ■

IV. Conclusion

The well-know push-sum based subgradient algorithm for distributed convex optimization over unbalanced directed graphs has been revisited. A novel analysis tool has been proposed, which improves the convergence rate of the subgradient-push algorithm from $O(\ln t/\sqrt{t})$ to $O(1/\sqrt{t})$, which is the same as that of the single-agent subgradient method and thus optimal. As a future work, the proposed tool is expected to be applicable to analyze other push-sum based algorithms and improve/simplify their convergence analyses, for example, DEXTRA [15] and Push-DIGing [17]. Another future direction is to extend the proposal tool to push-sum based distributed algorithms with communication delays and asynchronous updating.

References

[1] M. Cao, A.S. Morse, and B.D.O. Anderson. Reaching a consensus in a dynamically changing environment: A graphical approach. *SIAM Journal on Control and Optimization*, 47(2):575–600, 2008.

[2] L. Xiao and S. Boyd. Fast linear iterations for distributed averaging. *Systems & Control Letters*, 53(1):65–78, 2004.

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

[4] K.I. Tsianos, S. Lawlor, and M.G. Rabbat. Push-sum distributed dual averaging for convex optimization. In *Proceedings of the 51st IEEE Conference on Decision and Control*, pages 5453–5458, 2012.

[5] A. Nedić and A. Olshevsky. Distributed optimization over time-varying directed graphs. *IEEE Transactions on Automatic Control*, 60(3):601–615, 2015.

[6] A. Nedić and A. Olshevsky. Stochastic gradient-push for strongly convex functions on time-varying directed graphs. *IEEE Transactions on Automatic Control*, 61(12):3936–3947, 2016.

[7] M. Assran, N. Loizou, N. Ballas, and M. Rabbat. Stochastic gradient push for distributed deep learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 344–353, 2019.

[8] Y. Lin, K. Zhang, Z. Yang, Z. Wang, T. Başar, R. Sandhu, and J. Liu. A communication-efficient multi-agent actor-critic algorithm for distributed reinforcement learning. In *Proceedings of the 58th IEEE Conference on Decision and Control*, pages 5562–5567, 2019.

[9] J. Zhu and J. Liu. A distributed algorithm for multi-armed bandit with homogeneous rewards over directed graphs. In *Proceedings of the 2021 American Control Conference*, pages 3038–3043, 2021.

[10] Y. Lin, V. Gupta, and J. Liu. Finite-time error bounds for distributed linear stochastic approximation. 2021. arXiv:2111.12665 [cs.LG].

[11] F. Bénézit, V. Blondel, P. Thiran, J. N. Tsitsiklis, and M. Vetterli. Weighted gossip: distributed averaging using non-doubly stochastic matrices. In *Proceedings of the 2010 IEEE International Symposium on Information Theory*, pages 1753–1757, 2010.

[12] C.N. Hadjicostis and T. Charalambous. Average consensus in the presence of delays in directed graph topologies. *IEEE Transactions on Automatic Control*, 59(3):763–768, 2013.

[13] B. Gerencsér and L. Gerencsér. Tight bounds on the convergence rate of generalized ratio consensus algorithms. *IEEE Transactions on Automatic Control*, 67(4):1669–1684, 2021.

[14] J. Liu and A.S. Morse. Asynchronous distributed averaging using double linear iterations. In *Proceedings of the 2012 American Control Conference*, pages 6620–6625, 2012.

[15] C. Xi and U.A. Khan. DEXTRA: A fast algorithm for optimization over directed graphs. *IEEE Transactions on Automatic Control*, 62(10):4980–4993, 2017.

[16] W. Shi, Q. Ling, G. Wu, and W. Yin. EXTRA: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.

[17] A. Nedić, A. Olshevsky, and W. Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017.

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

[19] T. Yang, X. Yi, J. Wu, Y. Yuan, D. Wu, Z. Meng, Y. Hong, H. Wang, Z. Lin, and K.H Johansson. A survey of distributed optimization. *Annual Reviews in Control*, 47:278–305, 2019.

[20] A. Nedić and J. Liu. Distributed optimization for control. *Annual Review of Control, Robotics, and Autonomous Systems*, 1:77–103, 2018.

[21] D.K. Molzahn, F. Dörfler, H. Sandberg, S.H. Low, S. Chakrabarti, R. Baldick, and J. Lavaei. A survey of distributed optimization and control algorithms for electric power systems. *IEEE Transactions on Smart Grid*, 8(6):2941–2962, 2017.

[22] B. Gharesifard and J. Cortés. Distributed continuous-time convex optimization on weight-balanced digraphs. *IEEE Transactions on Automatic Control*, 59(3):781–786, 2013.

[23] A. Nedić, A. Olshevsky, and M. G Rabbat. Network topology and communication-computation tradeoffs in decentralized optimization. *Proceedings of the IEEE*, 106(5):953–976, 2018.

[24] B. Polyak. A general method for solving extremum problems. *Doklady Akademii Nauk*, 8(3):593–597, 1967.

[25] A. Nedić, A. Olshevsky, A. Ozdaglar, and J.N. Tsitsiklis. On distributed averaging algorithms and quantization effects. *IEEE Transactions on Automatic Control*, 54(11):2506–2517, 2009.

[26] Y. Lin and J. Liu. Push-subgradient is of the optimal convergence rate. *arXiv preprint*, 2022. arXiv:2203.16623 [math.OC].

[27] F. Iutzeler, P. Ciblat, and W. Hachem. Analysis of sum-weight-like algorithms for averaging in wireless sensor networks. *IEEE Transactions on Signal Processing*, 61(11):2802–2814, 2013.

[28] B. Touri. *Product of Random Stochastic Matrices and Distributed Averaging*. Springer Science & Business Media, 2012.

[29] A. Nedić and J. Liu. On convergence rate of weighted-averaging dynamics for consensus problems. *IEEE Transactions on Automatic Control*, 62(2):766–781, 2017.

[30] F. Saadatniaki, R. Xin, and U.A. Khan. Decentralized optimization over time-varying directed graphs with row and column-stochastic matrices. *IEEE Transactions on Automatic Control*, 65(11):4769–4780, 2020.

[31] A. Kolmogoroff. Zur theorie der markoffschen ketten. *Mathematische Annalen*, 112(1):155–160, 1936.

[32] D. Blackwell. Finite non-homogeneous chains. *Annals of Mathematics*, 46(4):594–599, 1945.