Push-Subgradient 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 push-subgradient algorithm actually converges at a rate of $O(1/t)$, 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]. 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] and distributed reinforcement learning [5].

The push-sum algorithm was first proposed in [3] and sometimes also called weighted gossip [6], ratio consensus [7], and double linear iteration [8]. Although the analysis of the push-sum algorithm is elegant, the analyses of push-sum based algorithms are often quite complicated, e.g., push-subgradient [4], DEXTRA [9] (a push-sum based variant of the well-known EXTRA algorithm [10]) and Push-DIGing [11]. Actually, all these push-sum based algorithms rely on the pioneering analysis and results in [4].

Distributed optimization originated from the work of [12] and has achieved great success in both theory and practice; see survey papers [13]–[15]. 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 [16]), 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 [17].

The push-sum based subgradient algorithm proposed in [4] is the first distributed convex optimization algorithm which works for unbalanced directed graphs. There are two “gaps” in the analysis in [4]. 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 [4]), which is “unusual” compared with non-push-sum based distributed optimization algorithms (see e.g. [12]). Second, more importantly, the convergence rate derived in [4] is of order $O(\ln t/\sqrt{t})$, which is slower than that of the single-agent subgradient method, $O(1/t)$ (see Theorem 7 in [17]). 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. PUSH-SUBGRADIENT

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 $\mathcal{G}(t) = (\mathcal{V}, \mathcal{E}(t))$ 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 $\mathcal{G}(t)$ 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
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 $Z$, is nonempty.

Since each $f_i$ is not necessarily differentiable, the gradient descent method may not be applicable. Instead, the subgradient method [18] can be applied. For a convex function $h : \mathbb{R}^d \rightarrow \mathbb{R}$, a vector $g \in \mathbb{R}^d$ is called a subgradient of $h$ at point $x$ if

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

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,

$$h(y) - h(x) \geq -G\|y - x\|_2, \quad (2)$$

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 [18] and the first distributed subgradient method was proposed in [12], which is based on average consensus. The push-subgradient algorithm, proposed in [4], is as follows

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

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

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 $j \in \mathcal{N}(t)$. For all $i \in \mathcal{V}$ and $t$, $\sum_{j \in \mathcal{N}^- (t)} w_{ji}(t) = 1$.

A typical choice of $w_{ij}(t)$ is $1/|\mathcal{N}^- (t)|$ for all $j \in \mathcal{N}(t)$ 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 $ij$th entry equals $w_{ij}(t)$ if $j \in \mathcal{N}(t)$ and zero otherwise; in other words, we set $w_{ij}(t) = 0$ for all $j \notin \mathcal{N}(t)$. From Assumption 1 each $W(t)$ is a column stochastic matrix that is compliant with the neighbor graph $G(t)$. 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 push-subgradient 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 $\sum_{t=0}^{\infty} \alpha^2(t) < \infty$.

---

3A directed graph is strongly connected if it has a directed path from any vertex to any other vertex. The union of two directed graphs, $G_p$ and $G_q$, with the same vertex set, written $G_p \cup G_q$, is meant the directed graph with the same vertex set and edge set being the union of the edge set of $G_p$ and $G_q$. Since this union is a commutative and associative binary operation, the definition extends unambiguously to any finite sequence of directed graphs with the same vertex set.
and provides convergence rates for each individual agent.

uniformly bounded above by a positive number

Here \( \alpha \) are positive constants which satisfy

\[
(1 - \frac{1}{n^2}) = 1
\]

and that

\[
\lim_{t \to \infty} f \left( \frac{\sum_{\tau=0}^{T-1} \alpha(\tau) z_k(\tau)}{T} \right) = f(z^*)
\]

with a fixed stepsize \( \alpha(t) = 1/\sqrt{T} \) for \( T > 0 \) steps, i.e., \( t \in \{0, 1, \ldots, T - 1\} \), then

\[
f \left( \frac{\sum_{\tau=0}^{T-1} \alpha(\tau) z_k(\tau)}{T} \right) - f(z^*)
\]

steps, i.e., \( t \in \{0, 1, \ldots, T - 1\} \), then for any \( k \in \mathcal{V} \),

\[
\leq \frac{\|\bar{z}(0) - z^*\|_2^2 + G^2}{nT} + \frac{32nG^2}{\eta(1 - \mu)T} \left\| \sum_{i=1}^{n} x_i(0) + \frac{1}{\sqrt{T}} g_i(0) \right\|_2.
\]

Here the positive constants \( \eta \) and \( \mu \leq 1 \) 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

\[
\eta(1 - \mu) \sqrt{T}
\]

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 \( \{G_i\} \) 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 \( k \in \mathcal{V} \),

\[
f \left( \frac{\sum_{\tau=0}^{T-1} \alpha(\tau) z_k(\tau)}{T} \right) - f(z^*)
\]

\[
\leq \frac{\|\bar{z}(0) - z^*\|_2^2 + G^2}{nT} \sum_{\tau=0}^{T-1} \alpha(\tau) + \frac{2G\alpha(0)}{n} \sum_{i=1}^{n} \frac{\|\bar{z}(0) - z_i(0)\|_2}{\|z_i(0)\|_2} + \frac{32G^2}{\eta(1 - \mu)T} \left\| \sum_{i=1}^{n} x_i(0) + \frac{1}{\sqrt{T}} g_i(0) \right\|_2.
\]

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 \( k \in \mathcal{V} \),

\[
f \left( \frac{\sum_{\tau=0}^{T-1} \alpha(\tau) z_k(\tau)}{T} \right) - f(z^*)
\]

\[
\leq \frac{\|\bar{z}(0) - z^*\|_2^2 + G^2}{2nT} \sum_{\tau=0}^{T-1} \alpha(\tau) + \frac{G}{n} \sum_{i=1}^{n} \|\bar{z}(0) - z_i(0)\|_2 + \|z_k(0) - z_i(0)\|_2
\]

\[
+ \frac{32G^2}{\eta(1 - \mu)T} \left\| \sum_{i=1}^{n} x_i(0) + \frac{1}{\sqrt{T}} g_i(0) \right\|_2.\tag{8}
\]

III. ANALYSIS

In this section, we provide a novel analysis of the push-subgradient algorithm [4–4] 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, \( x_i(t) \) and \( y_i(t) \), which are updated as follows:

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

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

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

Let \( x(t) \) and \( y(t) \) be the vectors in \( \mathbb{R}^n \) whose \( i \)-th entries are \( x_i(t) \) and \( y_i(t) \), respectively. 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

\[
\sum_{t=1}^{n} x_i(t) = \sum_{t=1}^{n} x_i(0) \quad \text{and} \quad \sum_{t=1}^{n} y_i(t) = \sum_{t=1}^{n} y_i(0) = n \quad \text{for all} \ t \geq 0.
\]

**Lemma 1:** Suppose that \( \{G(t)\} \) is uniformly strongly connected. Then, for any fixed \( \tau \geq 0 \), \( W(t) \cdots W(\tau + 1)W(\tau) \) will converge to the set \( \{v^\top v : v \in \mathbb{R}^n, 1^\top v = 1, v > 0\} \) exponentially fast as \( t \to \infty \).\footnote{We use \( 0 \) and \( 1 \) 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 \( v > 0 \) to denote a positive vector, i.e., each entry of \( v \) is positive.}
The lemma is essentially the same as Corollary 2 (a) in [4]. Suppose \( \{G_t\} \) 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 \( \{v(t)\} \) such that for all \( i, j \in \mathcal{V} \) and \( t \geq \tau \geq 0 \),

\[
|W(t) \cdots W(\tau + 1)W(\tau)|_{ij} - v(t) \leq c\mu^{t-\tau}, \tag{11}
\]

where \( |\cdot|_{ij} \) denotes the \( ij \)th entry of a matrix. In [4], it has been shown that \( c = 4 \) and \( \mu = (1 - \frac{3}{\sqrt{n}})^{1/k} \).

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

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

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 \( \{G(t)\} \) 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 [4], which further proves that if \( \{G_t\} \) is uniformly strongly connected by sub-sequences of length \( L \), then \( \eta \geq \frac{1}{\sqrt{n}e^2} \).

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

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

which implies that \( z(t + 1) = S(t)z(t) \) where \( z(t) \) is the vector in \( \mathbb{R}^n \) whose \( i \)th entry is \( z_i(t) \). 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 \( t > \tau \geq 0 \), holds \( \Phi_S(t, \tau)z_j(t) = \Phi_W(t, \tau)z_j(t) \).

**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), \( \Phi_S(\tau + 1, \tau)z_j(\tau) = s_{ij}(\tau) = \frac{y_i(\tau)}{y_i(\tau + 1)} \Phi_W(\tau + 1, \tau)z_j(\tau) \). 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 \( t = h + 1 \).

\[
\Phi_S(t + 1, \tau)z_j(t) = \Phi_W(t + 1, \tau)z_j(t) = y_j(\tau)\frac{y_i(t)}{y_i(t + 1)} = \frac{y_j(\tau)y_i(t)}{y_i(t + 1)} = \frac{n y_j(\tau)|\Phi_W(t + 1, \tau)z_j(t) - y_j(\tau)|}{ny_i(t + 1)}.
\]

Then,

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

which establishes the claim by induction.

More can be said.

**Lemma 4:** Suppose that \( \{G(t)\} \) is uniformly strongly connected. Then, for any fixed \( \tau \geq 0 \), \( S(t) \cdots S(\tau + 1)S(\tau) \) will converge to \( \frac{1}{y_i(t)} \Phi_W(\tau) \).

**Proof of Lemma 4** From Lemma 1 for any given \( \tau \geq 0 \), there holds \( \lim_{t \to \infty} [\Phi_W(t, \tau)] = v(\tau, \infty)I^T \), 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

\[
\lim_{t \to \infty} [\Phi_S(t, \tau)]_{ij} = \lim_{t \to \infty} \frac{y_j(\tau)}{y_i(t)}[\Phi_W(t, \tau)]_{ij} = \lim_{t \to \infty} \frac{y_j(\tau)}{y_i(t)} \lim_{t \to \infty} \frac{y_i(t)}{v_i(\tau, \infty)}y_k(\tau) = \lim_{t \to \infty} \frac{y_j(\tau)}{v_i(\tau, \infty)}y_k(\tau) = \frac{1}{v_i(\tau, \infty)}y_k(\tau),
\]

where in (a) we used the fact that \( v(\tau, \infty) > 0 \) 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 \( \{G(t)\} \) is uniformly strongly connected. Then, for any fixed \( \tau \geq 0 \), \( S(t) \cdots S(\tau + 1)S(\tau) \) will converge to \( \frac{1}{y_i(t)} \Phi_W(\tau) \) exponentially fast as \( t \to \infty \).

**Proof of Proposition 1** From (13), there exist constants \( c > 0 \) and \( \mu \in [0, 1) \) and a sequence of stochastic vectors \( \{v(t)\} \) such that \( |[\Phi_W(t + 1, \tau)]_{ij} - v(t)| \leq c\mu^{t-\tau} \) for all \( i, j \in \mathcal{V} \) and \( t \geq \tau \geq 0 \). Recall that \( \sum_{i=1}^{n} y_i(t) \) always equals \( n \) and, by Lemma 2 all \( y_i(t) \) are always positive. From Lemma 3 for all \( t \geq \tau \geq 0 \),

\[
\frac{|\Phi_S(t + 1, \tau)z_j(t) - y_j(\tau)|}{ny_i(t + 1)} = \frac{|y_j(\tau)(\Phi_W(t + 1, \tau)z_j(t) - y_j(\tau)|}{ny_i(t + 1)} = \frac{ny_j(\tau)|\Phi_W(t + 1, \tau)z_j(t) - y_j(\tau)|}{ny_i(t + 1)} = \frac{ny_j(\tau)(\Phi_W(t + 1, \tau)z_j(t) - v_i(t) + v_i(t))}{ny_i(t + 1)} = \frac{ny_j(\tau)(\Phi_W(t + 1, \tau)z_j(t) - v_i(t) + v_i(t))}{ny_i(t + 1)}.
\]
which implies that
\[
\left| \Phi_{S(t+1, \tau)} \right|_{ij} - \frac{y_j(\tau)}{n} \leq \frac{n y_j(\tau) \left| \Phi_{W(t+1, \tau)} \right|_{ij} - v_i(t)}{n y_i(t+1)} + \frac{y_j(\tau) \sum_{k=1}^{n} \left| \Phi_{W(t+1, \tau)} \right|_{ik} - v_i(t) y_k(\tau)}{n y_i(t+1)} + \frac{ny_j(\tau) + y_j(\tau) \sum_{k=1}^{n} \left| \Phi_{W(t+1, \tau)} \right|_{ik} - v_i(t) y_k(\tau)}{n y_i(t+1)} + \frac{2y_j(\tau) \mu^{t-\tau}}{y_i(t+1)} \leq \frac{2\eta \mu^{t-\tau}}{\eta \mu^{t-\tau}},
\]
where we used Lemma 2 in the last inequality. The above immediately implies the proposition.

The proposition immediately implies the following results.

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

**Proof of Corollary 2**: The corollary is a special case of Proposition 1 by setting \( \tau = 0 \).

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

**Proof of Corollary 2**: From (13), \( z(t+1) = S(t) z(t) = S(t) \cdots S(0) z(0) = S(t) \cdots S(0) x(0) \). From Corollary 1, \( z(t+1) \) will converge to \( \frac{1}{\mu} \mathbf{1} \mathbf{1}^\top x(0) = \left( \frac{1}{n} \sum_{i=1}^{n} x_i(0) \right) \mathbf{1} \) exponentially fast as \( t \to \infty \), which completes the proof.

Although the above proof of Corollary 2 looks more complicated than the conventional convergence proof of the push-sum algorithm (e.g., [3], [7], [8]), 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 \( i, j \in V \) and \( t \), \( s_{ij}(t) \geq \gamma \) whenever \( s_{ij}(t) > 0 \). For all \( i \in V \) and \( t \), \( \sum_{j=1}^{n} s_{ij}(t) = 1 \).

**Lemma 5**: Suppose that Assumption 3 holds. Then, \( s_{ij}(t) \) satisfies Assumption 1 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 \( w_{ij}(t) > 0 \),
\[
s_{ij}(t) = \frac{w_{ij}(t)y_j(t)}{\sum_{k=1}^{n} w_{ik}(t)y_k(t)} \geq \frac{\beta \eta}{n} \sum_{k=1}^{n} \frac{w_{ik}(t)y_k(t)}{w_{ik}(t)y_k(t)} = \frac{\beta \eta}{n}.
\]

The above inequality and Assumption 1 imply that \( s_{ij}(t) \) satisfies the first sentence of Assumption 3 with \( \gamma = \beta \eta/n \). For the second sentence of Assumption 3 it is easy to see 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 \( i \in V \) 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 \( G(t) \).

**Lemma 6**: The graph of \( S(t) \) is the same as the graph of \( W(t) \) for all \( t \).

**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 [20]. To analyze such a process, we appeal to the following concept. 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 [21]. It was shown by Blackwell [22] 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” it has a unique absolute probability sequence [23, 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 [24]) Let \( \{S(t)\} \) be a sequence of stochastic matrices satisfying Assumption 3. If the graph sequence of \( \{G(t)\} \) is uniformly strongly connected, then there exists a unique absolute probability sequence \( \{\pi(t)\} \) for the matrix sequence \( \{S(t)\} \) and a constant \( \pi_{\text{min}}(0,1) \) such that \( \pi_i(t) \geq \pi_{\text{min}} \) for all \( i \) and \( t \).

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

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

The proposition is a consequence of Lemma 1 in [23]. 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 \( \tau \geq 0 \)
\[ \pi^T(\tau) = \pi^T(\tau + 1)S(\tau) = \pi^T(\tau + 2)S(\tau + 1)S(\tau) \]
\[ = \lim_{t \to \infty} \pi^T(\tau + 1)S(t) \cdots S(\tau + 1)S(\tau) \]
\[ = \lim_{t \to \infty} \frac{1}{n} \pi^T(\tau + 1)1y^T(\tau) = \frac{1}{n} y^T(\tau), \]
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 \( \pi^T(t + 1)S(t) \). To see this, from (12) and Assumption 11 for \( j \in V \)
\[ [\pi^T(t + 1)S(t)]_j = \sum_{i=1}^{n} \frac{y_i(t + 1)}{n} s_{ij}(t) \]
\[ = \sum_{i=1}^{n} \frac{y_i(t + 1)}{n} w_{ij}(t)y_j(t) = \sum_{i=1}^{n} \frac{w_{ij}(t)y_j(t)}{n} \]
\[ = \frac{y_j(t)}{n} = \pi_j^T(t). \]
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 push-sum 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 push-subgradient 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 push-subgradient algorithm.

B. Push-Subgradient

We first rewrite the push-subgradient algorithm as follows. From (3)–(4), we have
\[ z_i(t + 1) = \frac{x_i(t + 1)}{y_i(t + 1)} = \sum_{j=1}^{n} \frac{w_{ij}(t)[x_j(t) - \alpha(t)g_j(t)]}{\sum_{j=1}^{n} w_{ij}(t)y_j(t)} \]
\[ = \sum_{j=1}^{n} \frac{w_{ij}(t)[x_j(t) - \alpha(t)g_j(t)]}{\sum_{k=1}^{n} w_{ik}(t)y_k(t)} \]
\[ = \sum_{j=1}^{n} \frac{[w_{ij}(t)y_j(t)]}{[\sum_{k=1}^{n} w_{ik}(t)y_k(t)]} \left[ z_j(t) - \alpha(t)g_j(t) \right] y_j(t) \]
\[ = \sum_{j=1}^{n} s_{ij}(t) \left[ z_j(t) - \alpha(t)g_j(t) \right] y_j(t), \]
where \( s_{ij}(t) \) is defined in (12). In addition,
\[ \bar{z}(t + 1) = \frac{1}{n} \sum_{i=1}^{n} z_i(t + 1) \]
\[ = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} s_{ij}(t) \left[ z_j(t) - \alpha(t)g_j(t) \right] y_j(t). \]
Define a time-varying Lyapunov function
\[ \langle z(t) \rangle = \pi^T(t)z(t). \]
Then, from Definition 2 we have
\[ \langle z(t + 1) \rangle = \sum_{i=1}^{n} \pi_i(t + 1)z_i(t + 1) \]
\[ = \sum_{i=1}^{n} \sum_{j=1}^{n} \pi_i(t + 1)s_{ij}(t) \left[ z_j(t) - \alpha(t)g_j(t) \right] y_j(t) \]
\[ = \sum_{j=1}^{n} \pi_j(t) \left[ z_j(t) - \alpha(t)g_j(t) \right] = \langle z(t) \rangle - \alpha(t) \sum_{i=1}^{n} g_i(t), \]
where we use the Proposition 2 in the last equality.

To prove Theorem 1 we need the following lemma.

Lemma 8: Suppose that \( \{ G_t \} \) 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 \( i \in V \),
\[ \| z_i(t + 1) - \frac{1}{n} \sum_{j=1}^{n} (x_j(t) - \alpha(t)g_j(t)) \|_2 \]
\[ \leq \frac{8}{\eta} \mu^2 \left( \sum_{i=1}^{n} x_i(0) + \alpha(0)g_i(0) \right) + \frac{8nG}{\eta} \sum_{s=1}^{t} \mu^{t-s} \alpha(s). \]
Suppose, in addition, that Assumption 2 holds. Then, for all \( t \geq 0 \) and \( i \in V \),
\[ \| z_i(t + 1) - \frac{1}{n} \sum_{j=1}^{n} (x_j(t) - \alpha(t)g_j(t)) \|_2 \]
\[ \leq \frac{8}{\eta} \mu^2 \left( \sum_{i=1}^{n} x_i(0) + \alpha(0)g_i(0) \right) + \frac{8nG}{\eta(1 - \mu)} \left( \alpha(0) \mu^{t/2} + \alpha(t)^{t/2} \right). \]
Here \( \eta > 0 \) and \( \mu \in (0, 1) \) are defined in Lemma 2 and (11), respectively.

Proof of Lemma 8: Let \( h(t) = x(t) - \alpha(t)g(t) \). Then,
\[ h(t + 1) = x(t + 1) - \alpha(t + 1)g(t + 1) \]
\[ = W(t)h(t) - \alpha(t + 1)g(t + 1) \]
\[ = \Phi_W(t + 1, 0)h(0) - \sum_{l=1}^{t} \alpha(l)\Phi_W(t + 1, l)g(l) - \alpha(t + 1)g(t + 1). \]
In addition,

\[
W(t + 1)h(t + 1) = \Phi_W(t + 2, 0)h(0) - \sum_{l=1}^{t+1} \alpha(l)\Phi_W(t + 2, l)g(l) \tag{14}
\]

and

\[
1^T h(t + 1) = 1^T h(0) - \sum_{l=1}^{t+1} \alpha(l) 1^T g(l). \tag{15}
\]

From Lemma \[\text{Lemma 1}\] and \eqref{11}, there exists a sequence \{\phi(t)\} of stochastic vectors, such that for all \(i, j \in \mathcal{V}\) and \(t \geq s \geq 0\)

\[
\left[\Phi_W(t + 1, s)\right]_{ij} - \phi_i(t) \leq 4\mu^{t-s}.
\]

Let \(D(s : t) = \Phi_W(t + 1, s) - \phi(t)1^T\). From \eqref{14} and \eqref{15},

\[
W(t + 1)h(t + 1) - \phi(t + 1)1^T h(t + 1) = \Phi_W(t + 2, 0)h(0) - \sum_{l=1}^{t+1} \alpha(l)\Phi_W(t + 2, l)g(l)
\]

\[
- \phi(t + 1)(1^T h(0) - \sum_{l=1}^{t+1} \alpha(l) 1^T g(l))
\]

\[
= (\Phi_W(t + 2, 0) - \phi(t + 1)1^T)h(0)
\]

\[
- \sum_{l=1}^{t+1} \alpha(l)\Phi_W(t + 2, l) - \phi(t + 1)1^T g(l)
\]

\[
= D(0 : t + 1)h(0) - \sum_{l=1}^{t+1} \alpha(l)D(l : t + 1)g(l),
\]

which implies that

\[
x(t + 1) = W(t)h(t)
\]

\[
= \phi(t)1^T h(t) + D(0 : t)h(0) - \sum_{l=1}^{t} \alpha(l)D(l : t)g(l).
\]

Moreover, \(y(t + 1) = \Phi_W(t + 1, 0)g(0) = D(0 : t)1 + \phi(t)\). Thus, for all \(i \in \mathcal{V}\)

\[
z_i(t + 1) = \frac{g(t)1}{n} = \frac{x_i(t + 1) - \sum_{j=1}^{n} (x_j(t) - \alpha(t)g_j(t))}{n}
\]

\[
= \frac{\phi_i(t)h(t)1 + \sum_{k=1}^{n}[D(0 : t)_{ij}h_k(0)]}{[D(0 : t)_{ij} + \phi_i(t)]} + \frac{1^T h(t)}{n}
\]

\[
- \frac{\sum_{l=1}^{t} \alpha(l)\sum_{k=1}^{n}[D(l : t + 1)_{ij}h_k(l)]}{[D(0 : t)_{ij} + \phi_i(t)]} - \frac{h(t)1}{n}
\]

\[
= \frac{n\sum_{k=1}^{n}[D(0 : t)_{ij}h_k(0)] - [D(0 : t)_{ij}1]h(t)}{n[D(0 : t)_{ij} + \phi_i(t)]}
\]

\[
- \frac{n\sum_{l=1}^{t} \alpha(l)\sum_{k=1}^{n}[D(l : t + 1)_{ij}h_k(l)]}{n[D(0 : t)_{ij} + \phi_i(t)]}.
\]

From the definition of \(D(0 : t)\), we have \([D(0 : t)1]_i + n\phi_i(t) = \Phi_W(t + 1, 0)1_i\),

\[
n\phi_i(t) = \Phi_W(t + 1, 0)1_i \geq \eta. \text{ Therefore,}
\]

\[
\|z_i(t + 1) - \frac{h(t)1}{n}\|_2 \leq \frac{1}{\eta} \left[n\sum_{j=1}^{n} (x_j(t) - \alpha(t)g_j(t))\right]_2 + \frac{\|D(0 : t)1\|_2 + \|h(t)1\|_2}{n[D(0 : t)_{ij} + \phi_i(t)]}
\]

\[
+ \frac{n\sum_{l=1}^{t} \alpha(l)\sum_{k=1}^{n}[D(l : t)_{ij}h_k(l)]}{n[D(0 : t)_{ij} + \phi_i(t)]}
\]

\[
\leq \frac{1}{\eta} \left[4\mu^t\|\sum_{k=1}^{n} h_k(0)\|_2 + \sum_{l=1}^{t} \alpha(l)4\mu^{t-l}\|\sum_{k=1}^{n} g_k(l)\|_2 + 4\mu^t\|h(t)1\|_2\right]
\]

\[
\leq \frac{1}{\eta} \left[4\mu^t\|\sum_{k=1}^{n} h_k(0)\|_2 + \sum_{l=1}^{t} \alpha(l)\|\sum_{k=1}^{n} g_k(l)\|_2 + \sum_{l=1}^{t} \alpha(l)\|\sum_{k=1}^{n} g_k(l)\|_2 + 4\mu^t\|h(t)1\|_2\right]
\]

In addition, from \eqref{15}, we have \(\|1^T h(t + 1)\|_2 \leq \|1^T h(0)\|_2 + \|\sum_{l=1}^{t+1} \alpha(l)1^T g(l)\|_2\). Then,

\[
\|z_i(t + 1) - \frac{h(t)1}{n}\|_2 \leq \frac{4\mu^t\|\sum_{k=1}^{n} h_k(0)\|_2 + \sum_{l=1}^{t} \alpha(l)\mu^{t-l}\|\sum_{k=1}^{n} g_k(l)\|_2 + \mu^t\|1^T h(t)1\|_2 + \|\sum_{l=1}^{t} \alpha(l)1^T g(l)\|_2}{\eta}
\]

\[
\leq \frac{8\mu^t\|\sum_{k=1}^{n} h_k(0)\|_2 + 2\sum_{l=0}^{t} \alpha(l)\mu^{t-l}\|\sum_{k=1}^{n} g_k(l)\|_2 + \sum_{l=0}^{t} \alpha(l)\|\sum_{k=1}^{n} g_k(l)\|_2}{\eta}
\]

\[
+ \|\sum_{l=0}^{t} \alpha(l)\mu^{t-l}\|\sum_{k=1}^{n} g_k(l)\|_2 \right]
\]

Then, for all \(t \geq 0\) and \(i \in \mathcal{V}\),

\[
\|z_i(t + 1) - \frac{1}{n} \sum_{j=1}^{n} (x_j(t) - \alpha(t)g_j(t))\|_2 \leq \frac{8\mu^t\|\sum_{i=1}^{n} x_i(0) + \alpha(0)g_i(0)\|_2 + \sum_{s=1}^{t} \mu^{t-s}\|\sum_{i=1}^{n} g_i(s)\|_2 + \frac{8nG}{\eta}\sum_{s=0}^{t} \mu^{t-s}\alpha(s)}{\eta}
\]
If the stepsize sequence \( \{\alpha(t)\} \) satisfies Assumption 2,

\[
\|z_t(t+1) - \frac{1}{n} \sum_{j=1}^{n} (x_j(t) - \alpha(t)g_j(t))\|_2 \\
\leq \frac{8}{\eta} \mu^t \| \sum_{i=1}^{n} x_i(0) + \alpha(0)g_i(0) \|_2 \\
+ \frac{8nG}{\eta} \left( \sum_{s=0}^{t} \alpha(s) + \sum_{s=1}^{t} \mu^{t-s} \alpha(s) \right) \\
\leq \frac{8}{\eta} \mu^t \| \sum_{i=1}^{n} x_i(0) + \alpha(0)g_i(0) \|_2 \\
+ \frac{8nG}{\eta(1-\mu)} \left( \alpha(0)\mu^{t/2} + \alpha(\lceil \frac{t}{2} \rceil) \right).
\]

This completes the proof.

We are now in a position to prove Theorem 2.

**Proof of Theorem 2.** From Lemma \( \mathcal{L} \) for all \( t \geq 0 \) and \( i \in \mathcal{V} \),

\[
\|z(t+1) - z_t(t+1)\|_2 + \|\tilde{z}(t+1) - z_t(t+1)\|_2 \\
\leq \|z(t+1)\|_2 - \frac{1}{n} \sum_{k=1}^{n} (x_k(t) - \alpha(t)g_k(t))\|_2 \\
+ \|\tilde{z}(t+1)\|_2 - \frac{1}{n} \sum_{k=1}^{n} (x_k(t) - \alpha(t)g_k(t))\|_2 \\
+ 2\|z_t(t+1) - z_t(t+1)\|_2,
\]

which implies that

\[
\|z(t+1) - z_t(t+1)\|_2 + \|\tilde{z}(t+1) - z_t(t+1)\|_2 \\
\leq 4\|z_t(t+1)\|_2 - \frac{1}{n} \sum_{k=1}^{n} (x_k(t) - \alpha(t)g_k(t))\|_2 \\
\leq \frac{32}{\eta} \mu^t \| \sum_{i=1}^{n} x_i(0) + \alpha(0)g_i(0) \|_2 + \frac{32nG}{\eta} \sum_{s=0}^{t} \mu^{t-s} \alpha(s).
\]

(16)

In addition, when the stepsize sequence \( \{\alpha(t)\} \) satisfies Assumption 2

\[
\|z(t+1) - z_t(t+1)\|_2 + \|\tilde{z}(t+1) - z_t(t+1)\|_2 \\
\leq \frac{32}{\eta} \mu^t \| \sum_{i=1}^{n} x_i(0) + \alpha(0)g_i(0) \|_2 \\
+ \frac{32nG}{\eta(1-\mu)} \left( \alpha(0)\mu^{t/2} + \alpha(\lceil \frac{t}{2} \rceil) \right).
\]

(17)

From the update of \( (z(t)) \), we have

\[
\|z(t+1) - z^*\|_2 = \|z(t) - z^* - \frac{\alpha(t)}{n} \sum_{i=1}^{n} g_i(t)\|_2 \\
\leq \|z(t) - z^*\|_2 + \frac{\alpha(t)}{n} \sum_{i=1}^{n} g_i(t)\|_2 \\
- 2\|z(t) - z^*\|_2 + \alpha^2(t)G^2 \\
- 2\|z(t) - z^*\|_2 \frac{\alpha(t)}{n} \sum_{i=1}^{n} g_i(t)).
\]

(18)

In addition,

\[
((z(t)) - z^*)^T g_t(t) \\
= ((z(t)) - z_t(t) + z_t(t) - z^*)^T g_t(t) \\
= ((z(t)) - z_t(t))^T g_t(t) + (z_t(t) - z^*)^T g_t(t) \\
\geq f_t(z_t(t)) - f_t(z^*) - G\|z_t(t) - z_t(t)\|_2 \\
\geq f_t(\tilde{z}(t)) - f_t(z^*) - G\|z_t(t) - z_t(t)\|_2 - G\|\tilde{z}(t) - z_t(t)\|_2,
\]

(19)

where we used (1) and (2) in deriving (19), and made use of (19) to get (20).

Combining (18) and (20), we have

\[
\|z(t+1)) - z^*\|_2^2 \\
\leq \|z(t)) - z^*\|_2^2 + \alpha^2(t)G^2 - 2\alpha(t)(f(\tilde{z}(t)) - f(z^*)) \\
+ \frac{2Ga(t)}{n} \sum_{i=1}^{n} \|z(t)) - z_t(t)\|_2 + \|\tilde{z}(t) - z_t(t)\|_2,
\]

which implies that

\[
2\alpha(t)(f(\tilde{z}(t)) - f(z^*)) \\
\leq \|z(t)) - z^*\|_2^2 + \alpha^2(t)G^2 - \|z(t+1)) - z^*\|_2^2 \\
+ \frac{2Ga(t)}{n} \sum_{i=1}^{n} \|z(t)) - z_t(t)\|_2 + \|\tilde{z}(t) - z_t(t)\|_2.
\]

Summing this up, we obtain

\[
\sum_{t=0}^{T} 2\alpha(t)(f(\tilde{z}(t)) - f(z^*)) \\
\leq \|z(0)) - z^*\|_2^2 + \|z(t+1)) - z^*\|_2^2 + \sum_{t=0}^{T} \alpha^2(t)G^2 \\
+ \sum_{t=0}^{T} \frac{2Ga(t)}{n} \sum_{i=1}^{n} \|z(t)) - z_t(t)\|_2 + \|\tilde{z}(t) - z_t(t)\|_2.
\]

In addition, since

\[
f \left( \frac{\sum_{t=0}^{T} \alpha(t)\tilde{z}(t)}{\sum_{t=0}^{T} \alpha(t)} \right) - f(z^*) \\
\leq \frac{\sum_{t=0}^{T} 2\alpha(t)(f(\tilde{z}(t)) - f(z^*))}{\sum_{t=0}^{T} 2\alpha(t)},
\]
then
\[
   f \left( \frac{\sum_{\tau=0}^{t} \alpha(\tau) \bar{z}(\tau)}{\sum_{\tau=0}^{t} \alpha(\tau)} \right) - f(z^*) 
   \leq \frac{\|\bar{z}(0) - z^*\|_2^2 - \|\bar{z}(t+1) - z^*\|_2^2 + \sum_{\tau=0}^{t} \alpha^2(\tau) G^2}{\sum_{\tau=0}^{t} 2\alpha(\tau)} 
\]

\[
   + \sum_{\tau=0}^{t} \frac{2G \alpha(\tau)}{n} \sum_{i=1}^{n} (\|\bar{z}(\tau)\|_2 - z_i(\tau))_2 + \|\bar{z}(\tau) - z_i(\tau)\|_2 \big) 
   \leq \frac{n \sum_{\tau=0}^{t} \alpha(\tau)}{\sum_{\tau=0}^{t} \alpha(\tau)} 
\]

\[
   + \frac{\|\bar{z}(0) - z^*\|_2^2 + 32G n^2}{\sum_{\tau=0}^{t} 2\alpha(\tau)}. \quad (21)
\]

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 (17) and (21), we have
\[
   f \left( \frac{\sum_{\tau=0}^{T-1} \alpha(\tau) \bar{z}(\tau)}{T} \right) - f(z^*) 
   \leq \frac{G \sum_{\tau=0}^{T-1} \|\bar{z}(\tau)\|_2 - z_i(\tau))_2 + \|\bar{z}(\tau) - z_i(\tau)\|_2 \big) }{nT} 
   + \frac{\|\bar{z}(0) - z^*\|_2^2 + 32G n^2}{2\sqrt{T}} 
\]

which implies that
\[
   f \left( \frac{\sum_{\tau=0}^{T-1} \alpha(\tau) \bar{z}(\tau)}{T} \right) - f(z^*) 
   \leq G \sum_{\tau=1}^{n} (\|\bar{z}(\tau)\|_2 + \|\bar{z}(\tau) - z_i(\tau)\|_2) 
   + \frac{\|\bar{z}(0) - z^*\|_2^2 + G^2}{2\sqrt{T}} + \frac{32nG^2}{\sqrt{T} \eta(1-\mu)} 
\]

\[
   + \frac{32G}{\eta(1-\mu)} \sum_{i=1}^{n} \sum_{\tau=0}^{T} g_i(\tau) \mu^\tau. 
\]

Since \( \|\bar{z}(0)\| = \frac{1}{n} \sum_{i=1}^{n} z_i(0) = z(0) \), we have derived (6).
We complete the proof.

2) If the stepsize \( \alpha(t) \) is fixed and \( \alpha(t) = 1/\sqrt{T} \) for all \( t \geq 0 \), then from (21), we have
\[
   f \left( \frac{\sum_{\tau=0}^{T-1} \bar{z}(\tau)}{T} \right) - f(z^*) 
   \leq \frac{\|\bar{z}(0) - z^*\|_2^2 + \sum_{\tau=0}^{T-1} \alpha^2(\tau) G^2}{nT} 
\]

\[
   + \frac{\|\bar{z}(0) - z^*\|_2^2 + 32G n^2}{2\sqrt{T}}. 
\]

Using (16), we have
\[
   f \left( \frac{\sum_{\tau=0}^{T-1} \bar{z}(\tau)}{T} \right) - f(z^*) 
   \leq \frac{\|\bar{z}(0) - z^*\|_2^2 + \|\bar{z}(0) - z_i(\tau)\|_2 \big) }{nT} 
\]

\[
   + \frac{\|\bar{z}(0) - z^*\|_2^2 + G^2}{2\sqrt{T}} + \frac{32nG^2}{T \eta} \sum_{\tau=0}^{T-2} \sum_{s=0}^{\tau} \mu^{\tau-s} \frac{1}{\sqrt{T}} 
\]

\[
   + \frac{32G}{T \eta} \sum_{i=1}^{n} \sum_{\tau=0}^{T} g_i(\tau) \mu^\tau, 
\]

it follows that
\[
   \lim_{t \to \infty} \frac{\sum_{\tau=0}^{t-1} \alpha(\tau) \mu^\tau}{\sum_{\tau=0}^{t-1} \alpha(\tau)} = 0 
\]

\[
   \lim_{t \to \infty} \frac{\sum_{\tau=0}^{t-1} \alpha(\tau) \mu^\tau}{\sum_{\tau=0}^{t-1} \alpha(\tau)} = 0. 
\]

From (5), we have
\[
   \lim_{t \to \infty} f \left( \frac{\sum_{\tau=0}^{t} \alpha(\tau) \bar{z}(\tau)}{\sum_{\tau=0}^{t} \alpha(\tau)} \right) - f(z^*) = 0. 
\]

2) If the stepsize is fixed and \( \alpha(t) = 1/\sqrt{T} \) for all \( t \geq 0 \), then from (5), we have
\[
   f \left( \frac{\sum_{\tau=0}^{T-1} \bar{z}(\tau)}{T} \right) - f(z^*) \leq O \left( \frac{1}{\sqrt{T}} \right). 
\]

This completes the proof.

We finally prove Theorem 3.
Proof of Theorem 2: From Lemma 8 for all $t \geq 0$ and $i, j \in \mathcal{V}$,
\[
\|z(t+1)\|_2^2 + \|z_j(t+1) - z_i(t+1)\|_2^2 \\
\leq \|z(t+1)\|_2^2 - \frac{1}{n} \sum_{k=1}^{n} \|x_k(t) - \alpha(t)g_k(t)\|_2^2 \\
+ \|z_j(t) - z_i(t)\|_2^2 + \frac{1}{n} \sum_{k=1}^{n} \|x_k(t) - \alpha(t)g_k(t)\|_2^2 \\
+ 2\|z_i(t+1) - \frac{1}{n} \sum_{k=1}^{n} \|x_k(t) - \alpha(t)g_k(t)\|_2^2 \\
\leq 4\|z_j(t+1) - \frac{1}{n} \sum_{k=1}^{n} \|x_k(t) - \alpha(t)g_k(t)\|_2^2 \\
\leq \frac{32}{\eta} \mu^2 \sum_{i=1}^{n} x_i(0) + \alpha(0)g_i(0) \|_2^2 + \frac{32nG}{\eta} \sum_{s=0}^{t} \mu^t \alpha(s).
\]
(22)

When the stepsize sequence $\{\alpha(t)\}$ satisfies Assumption 2,
\[
\|z(t+1) - z_i(t+1)\|_2^2 + \|z_j(t+1) - z_i(t+1)\|_2^2 \\
\leq \frac{32}{\eta} \mu^2 \sum_{i=1}^{n} x_i(0) + \alpha(0)g_i(0) \|_2^2 \\
+ \frac{32nG}{\eta(1-\mu)} \left( \alpha(0)\mu^{t/2} + \alpha(t) \right).
\]
(23)

From (1), (2) and (19), we have for any $k \in \mathcal{V}$
\[
\langle z(t) - z^* \rangle^T g_i(t) \\
\geq f_i(z_i(t)) - f_i(z_i^*) - G\|\langle z(t) - z_i(t) \rangle \|_2 \\
\geq f_i(z_k(t)) - f_i(z^*) - G\|\langle z(t) - z_i(t) \rangle \|_2 \\
- G\|z_k(t) - z_i(t)\|_2.
\]
(24)

Then, combining (13) and (24), we have
\[
\|z(t+1) - z^*\|_2^2 \\
\leq \|z(t) - z^*\|_2^2 + \alpha^2(t)G^2 - 2\alpha(t)(f(z_k(t)) - f(z^*)) \\
+ \frac{2Go(t)}{n} \sum_{i=1}^{n} \|\langle z(t) - z_i(t) \rangle \|_2 + \|z_k(t) - z_i(t)\|_2,
\]
which implies that
\[
2\alpha(t)(f(z_k(t)) - f(z^*)) \\
\leq \|z(t+1)\|_2^2 - \|z(t) - z^*\|_2^2 \\
+ \frac{2Go(t)}{n} \sum_{i=1}^{n} \|\langle z(t) - z_i(t) \rangle \|_2 + \|z_k(t) - z_i(t)\|_2.
\]

Summing this up, we obtain
\[
\sum_{\tau=0}^{t} 2\alpha(\tau)(f(z_k(\tau)) - f(z^*)) \\
\leq \|z(0) - z^*\|_2^2 + \|z(t+1)\|_2^2 + \sum_{\tau=0}^{t} \alpha^2(\tau)G^2 \\
+ \frac{2Go(t)}{n} \sum_{i=1}^{n} \|\langle z(\tau) - z_i(\tau) \rangle \|_2 + \|z_k(\tau) - z_i(\tau)\|_2.
\]

In addition, for $t \geq 0$ and $i, j \in \mathcal{V}$,
\[
f \left( \frac{\sum_{\tau=0}^{t} \alpha(\tau)z_k(\tau)}{\sum_{\tau=0}^{t} \alpha(\tau)} \right) - f(z^*) \\
\leq \sum_{\tau=0}^{t} 2\alpha(\tau)(f(z_k(\tau)) - f(z^*)) \\
\leq \|z(0) - z^*\|_2^2 + \|z(t+1)\|_2^2 + \sum_{\tau=0}^{t} \alpha^2(\tau)G^2 \\
+ \frac{2Go(t)}{n} \sum_{i=1}^{n} \|\langle z(\tau) - z_i(\tau) \rangle \|_2 + \|z_k(\tau) - z_i(\tau)\|_2 \\
\leq \sum_{\tau=0}^{t} \frac{2Go(\tau)}{n} \sum_{i=1}^{n} \|\langle z(\tau) - z_i(\tau) \rangle \|_2 + \|z_k(\tau) - z_i(\tau)\|_2 \\
+ \frac{\|z(0) - z^*\|_2^2 + \sum_{\tau=0}^{t} \alpha^2(\tau)G^2}{\sum_{\tau=0}^{t} \alpha(\tau)}.
\]
(25)

1) If the stepsize $\alpha(t)$ is time-varying and satisfies Assumption 2, then combining (23) and (25), we have
\[
f \left( \frac{\sum_{\tau=0}^{t} \alpha(\tau)z_k(\tau)}{\sum_{\tau=0}^{t} \alpha(\tau)} \right) - f(z^*) \\
\leq \|z(0) - z^*\|_2^2 + \sum_{\tau=0}^{t} \alpha^2(\tau)G^2 \\
+ \frac{Go(0) \sum_{i=1}^{n} \|\langle z(\tau) - z_i(\tau) \rangle \|_2 + \|z_k(\tau) - z_i(\tau)\|_2}{\sum_{\tau=0}^{t} \alpha(\tau)} \\
\leq 32G \sum_{i=1}^{n} \|\langle z(\tau) \rangle \|_2 + \|z_k(\tau) - z_i(\tau)\|_2 \\
&+ \frac{32Go(0) \sum_{i=1}^{n} \|\langle z(\tau) - z_i(\tau) \rangle \|_2 + \|z_k(\tau) - z_i(\tau)\|_2}{\sum_{\tau=0}^{t} \alpha(\tau)}.
\]

2) If the stepsize is fixed and $\alpha(t) = 1/\sqrt{T}$ for all $t \geq 0$, then from (25), we have
\[
f \left( \frac{\sum_{\tau=0}^{T-1} \frac{1}{\sqrt{T}}z_k(\tau)}{T} \right) - f(z^*) \\
\leq \frac{\|z(0) - z^*\|_2^2 + \sum_{\tau=0}^{T-1} \alpha^2(\tau)G^2}{2\sqrt{T}} \\
+ \frac{32Go(0) \sum_{i=1}^{n} \|\langle z(\tau) - z_i(\tau) \rangle \|_2 + \|z_k(\tau) - z_i(\tau)\|_2}{\sum_{\tau=0}^{T-1} \alpha(\tau)}.
\]

Using (22), we have
\[
f \left( \frac{\sum_{\tau=0}^{T-1} \frac{1}{\sqrt{T}}z_k(\tau)}{T} \right) - f(z^*) \\
\leq \frac{32G \sum_{i=1}^{n} \|\langle z(\tau) - z_i(\tau) \rangle \|_2 + \|z_k(\tau) - z_i(\tau)\|_2}{\sqrt{T}} \\
+ \frac{32Go(0) \sum_{i=1}^{n} \|\langle z(\tau) - z_i(\tau) \rangle \|_2 + \|z_k(\tau) - z_i(\tau)\|_2}{\sum_{\tau=0}^{T-1} \alpha(\tau)}.
\]
which implies that
\[
f\left(\frac{\sum_{\tau=0}^{T-1} z_k(\tau)}{T}\right) - f(z^*) \leq G \sum_{i=1}^{n} \|z(0) - z_i(0)\|_2 + \|z_k(0) - z_i(0)\|_2 \\
\quad + \frac{\|z(0) - z^*\|^2_2 + G^2}{2\sqrt{T}} + \frac{32nG^2}{T \eta (1 - \mu)} \\
\quad + \frac{32G}{T \eta (1 - \mu)} \sum_{i=1}^{n} x_i(0) + \frac{1}{\sqrt{T}} \eta_i(0)\|_2.
\]
Since \(\langle z(0)\rangle = \bar{z}(0)\), we have derived 8. This completes the proof.

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 push-subgradient 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 [9] and Push-DiGing [11]. 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] A. Nedić and A. Olshevsky. Distributed optimization over time-varying directed graphs. IEEE Transactions on Automatic Control, 60(3):601–615, 2015.
[5] 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.
[6] 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.
[7] C.N. Hadjicostis and T. Chalalambous. Average consensus in the presence of delays in directed graph topologies. IEEE Transactions on Automatic Control, 59(3):763–768, 2013.
[8] 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.
[9] 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.
[10] 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.
[11] A. Nedic, A. Olshevsky, and W. Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. SIAM Journal on Optimization, 27(4):2597–2633, 2017.