Geometrical Convergence Rate for Distributed Optimization with Time-Varying Directed Graphs and Uncoordinated Step-Sizes

Qingguo Lü, Huaqing Li*
College of Electronic and Information Engineering, Southwest University, Chongqing 400715
*Corresponding author: huaqingli@hotmail.com

Abstract: This paper studies a class of distributed optimization algorithms by a set of agents, where each agent only has access to its own local convex objective function, and the goal of the agents is to jointly minimize the sum of all the local functions. The communications among agents are described by a sequence of time-varying directed graphs which are assumed to be uniformly strongly connected. A column stochastic mixing matrices is employed in the algorithm which exactly steers all the agents to asymptotically converge to a global and consensual optimal solution even under the assumption that the step-sizes are uncoordinated. Two fairly standard conditions for achieving the geometrical convergence rate are established under the assumption that the objective functions are strong convexity and have Lipschitz continuous gradient. The theoretical analysis shows that the distributed algorithm is capable of driving the whole network to geometrically converge to an optimal solution of the convex optimization problem as long as the uncoordinated step-sizes do not exceed some upper bound. We also give an explicit analysis for the convergence rate of our algorithm through a different approach. Finally, simulation results illustrate the feasibility of the proposed algorithm and the effectiveness of the theoretical analysis throughout this paper.

Keywords: Distributed optimization, multi-agent systems, time-varying directed graphs, uncoordinated step-sizes, small gain theorem

I. INTRODUCTION

In recent years, with the rapid development of science and information technology, more and more researchers have paid their research attention to multi-agent systems and have obtained many remarkable achievements. On the one hand, multi-agent system provides a theoretical research method for modeling and analysis of complex systems. On the other hand, multi-agent system is also an important branch of distributed artificial intelligence research. As one of the most important research
subjects in the field of multi-agent systems, distributed optimization problem of multi-agent systems has attracted intensive research interest over the past few years due to their wide applications in distributed formation control of multiple autonomous vehicles [1], resource allocation in peer-to-peer communication networks [2], and distributed data fusion, information processing and decision making in wireless sensor networks [3]-[5], etc. Specifically, distributed optimization framework not only avoids establishing a long-distance communication system or data fusion center, but also provides better load balance to the network. In many networked systems, multi-agent systems are applied to solve such a distributed convex optimization problem where the goal is to minimize a sum of all the local objective functions, in which each local objective function is not known or shared by other agents.

In the existing literatures, distributed optimization problems are extensively solved by the (sub)gradient descent algorithm [6]-[7], the (sub)gradient-push descent algorithm [8], the fast (sub)gradient descent algorithm [9], and the dual averaging algorithm [10]. All of the above work [6]-[10] can be regarded as the consensus-based distributed (sub)gradient descent algorithms where each agent accomplishes a consensus step and then a descent step along the local (sub)gradient direction. During this time, many valuable consensus-based (sub)gradient algorithms have been addressed. Nedic et al. [8] show that the consensus-based distributed (sub)gradient descent (DGD) algorithm converges at a rate of $O(1/\sqrt{t})$ for convex Lipshitz and possibly nonsmooth objective functions when applying a diminishing step size. This coincides with the convergence rate of the centralized subgradient descent algorithm. Then, Nedic et al. [9] propose a (sub)gradient-push descent algorithm, which drives every agent to an optimal value under a fairly standard assumption of (sub)gradient boundedness. Noting that the (sub)gradient-push descent algorithm in [9] requires no knowledge of the number of agents or the graph sequence to implement. Theoretical analysis demonstrated that the (sub)gradient-push algorithm converges at a rate of $O(\ln t/\sqrt{t})$ when using a diminishing step size. This line of work has been extended to a variety of realistic conditions for distributed optimization, such as directed [8] or random communication graph [11], stochastic (sub)gradient errors [12], heterogeneous local constraints [13]-[14], linear scaling in network size [15]. Although these algorithms are intuitive and simple, it can not be neglected that they are usually slow. Among the reasons, even if the objective functions are differentiable and strongly convex, they still need to apply a diminishing step-size to converge to a consensus solution [8]-[23]. Also, the-abovementioned algorithms all require the assumption of bounded
(sub)gradient to achieve the exact optimal solution, which is another shortcoming. Furthermore, the abovementioned algorithms can be fast by using a fixed step-size, but they only converge to a neighborhood of the optimal solution set.

Recent work of Xu et al. [24] had coordinately put their sights on the analysis of a class of augmented distributed gradient method (Aug-DGM) of general linear time-invariant systems with fixed topology by applying the so-called Adapt–Then–Combine scheme. Specifically, this algorithm was proved to be convergent to an exact consensual minimizer for general convex and smooth objective functions when the uncoordinated step-size is sufficiently small. Then, Nedic et al. [25] studied the distributed optimization problems with coordinated step-size over time-varying undirected/directed graphs by combining the distributed inexact gradient method and the gradient tracking technique. The theoretical analysis showed that the distributed algorithm of Nedic et al. [25] was capable of driving the whole network to geometrically converge to an optimal solution under the assumption that the global objective function is strongly convex and has Lipschitz continuous gradient. Moreover, Nedic et al. in [26] also showed that the linear convergence rates could still be achieved in the studied algorithm for uncoordinated step-sizes, where the step-sizes do not exceed some upper bounds. The work of [26] could not be extended to directed multi-agent systems in terms of the proposed framework.

Based on the above discussions, the main focus of this paper is to establish a distributed optimization algorithm in general directed multi-agent systems, which can guarantee a consensus and geometrically converge to the optimum under uncoordinated yet bounded step-sizes. We look forward to facilitating the development of a generalized theory of distributed optimization, and our ultimate goal is to design more realistic step-sizes which are capable of adaptability and promoting the practical applications. More precisely, the main contributions of this paper can be summarized as follows: (i) A distributed algorithm is proposed to analyze the convex optimization problem of multi-agent systems over time-varying and yet uniformly strongly connected directed graphs. (ii) It is important to note that, unlike the distributed descent method or the push-sum protocol proposed in [6, 8], the proposed distributed algorithm takes account of applying uncoordinated step-sizes for local optimization and is exactly ensured to converge to the minimizer even with constant step-sizes. (iii) Theoretical analysis shows that the algorithm achieves a geometrical convergence rate as long as the uncoordinated step-sizes are smaller than an explicit upper bound and no positive lower bound is required when the objective functions are strong convexity and have Lipschitz continuous gradient. Specially, we construct linearly
convergent methods along with establishment of explicit bounds on their convergence rates. (iv)
Specifically, simulation results demonstrate that the algorithm has a faster convergence rates compared
with the well-known distributed (sub)gradient descent (DGD) algorithm [6] and the push-sum algorithm
[8].

The reminder of this paper is organized as follows. We begin in Section II where
we present the notations, formulate the problem of interest, introduce the communication network, and
give some useful assumptions. In Section III, we consider the optimization algorithm along with the
small gain theorem and some beneficial lemmas. The convergence and convergence rate results of the
proposed distributed optimization algorithm are established in Section IV. Furthermore, the
effectiveness of the algorithm is testified by applying a numerical example in Section V. Finally, some
conclusions and future directions are drawn in Section VI.

II. PROBLEM FORMULATION AND ASSUMPTIONS

A. Notations

Some standard notations throughout the paper are shown in the following. \( R, R^N \) \( R^{N\times N} \) refer to
the set of real numbers, the set of \( N \)-dimensional real column vectors and the set of \( N\times N \) real
matrices, respectively. \( I_N \) and \( 0_N \) denote the \( N \)-dimensional identity matrix and the \( N \)
dimensional zero matrix, respectively. We let \( 1_{N} \in R^{N} \) and \( 0_{N} \in R^{N} \) refer to the column vector with all
entries being one and zero, respectively. For a matrix \( W, W^T \) (probably a vector) and \( W^{-1} \) denote its
transpose and inverse, respectively. We use \( <\cdot,\cdot> \) to represent the inner product of two vectors. For a
vector \( x \in R^{N} \), we denote its average vector as \( \bar{x} = \frac{1}{N} 1_N^T x \) and its consensus violation as
\( \bar{x} = x - \frac{1}{N} 1_N 1_N^T x = (I-J)x = \bar{J}x \), where \( J = \frac{1}{N} 1_N 1_N^T \) and \( \bar{J} = I-J \) are two symmetric matrices. For a
vector \( x \in R^{N} \), we use \( \|x\| \) to denote the standard Euclidean norm, namely \( \|x\| = \sqrt{x^T x} \) and
\( \|x\|_\infty \) to denote the infinite norm. We also use \( \|x\|_j = \sqrt{<x,\bar{J}x>} \) to denote its \( \bar{J} \)
weighted (semi)-norm. Since \( \bar{J} = \bar{J}^T \bar{J} \), we have \( \|x\|_j = \|\bar{J}x\| \). We use \( \|W\| \) to denote the spectral norm of
matrix \( W \). \( \nabla f(x) : R^n \rightarrow R^n \) denotes the gradient of \( f(x) \). For a matrix \( W \), we write \( W_{ij} \) or
\( [W]_{ij} \) to denote its \( i, j \)’th entry. The notation \( \text{diag}(x) \) denotes a diagonal matrix whose entry in the \( i \)
-th row and -th column is , and the non-diagonal elements are zeros.

**B. Problem formulation**

In this paper, we consider a network of agents labeled by , which can only exchange information with each other via local communication. The target of the multi-agent group is to collectively solve the following convex optimization problem:

\[
\min f(x), \quad f(x) = \frac{1}{N} \sum_{i=1}^{N} f_i(x), \quad x \in \mathbb{R}^n
\]

over a global decision vector . For each agent , is the convex objective function. Assume that is privately known by agent , and probably different. Throughout the paper, we assume that the optimization problem (1) is solvable, namely, the optimal solution set is nonempty. We denote , the optimal value and an optimizer of the optimization problem (1). Specifically, the objective of this paper is to solve problem (1) in a distributed manner and require that the agents do not need to exchange the information of the objective function with each other, but only share their own states with their neighbors in each step.

**C. Communication Model**

The communication topology at time among agents can be modeled as a weighted directed graph , where is the set of vertices with representing th vertex, is the set of edges, and is the weighted adjacency matrix with . A directed edge denoted by a pair indicates that agent can arrive at agent or agent can receive information from agent . If an edge , then agent is called a neighbor of agent and . The in-neighbor set and out neighbor set of agent at time are denoted by and , respectively. The in-degree and out-degree of agent can be defined by and , respectively. A directed path from agent to agent is a sequence of edges in the directed graph with different nodes . A directed graph is strongly connected if and only if for any two distinct agents and in the set , there exists a directed path from agent to agent . We here adopt the following two standard assumptions on the network communication graphs, which are standard in the analysis of distributed optimization.
algorithm.

**Assumption 2.1:** The time-varying directed graph sequence \( G(k) \) is \( B_0 \)-strongly connected. That is, there exists an positive integer \( B_0 \) such that for any \( k \geq 0 \), the directed graph \( G(k) \) with vertices \( V \) and edge set \( E_{B_0}(k) = \bigcup_{s=k-B_0}^{k} E(s) \) is strongly connected.

**Assumption 2.2:** For any \( k = 0, 1, ..., \), the mixing matrix \( A(k) = [A_j(k)] \in \mathbb{R}^{N \times N} \) is defined by

\[
A_j(k) = \begin{cases}
1 & \text{whenever } j \in N_j^\text{out}(k) \\
0 & \text{otherwise}
\end{cases}
\]

where \( d_j^\text{out}(k) = |N_j^\text{out}(k)| \) is the out-degree of agent \( j \) at time \( k \). Also, we can conclude that \( A \) is a column stochastic matrix, i.e., \( \sum_{j=0}^{N} A_j(k) = 1 \) for any \( j \in V \).

Assumption 2.1 ensures that all agents can repeatedly interact each other through repeated communications with neighbors in the whole graph sequence \( G(k) \). Particularly, this assumption is considerably weaker than requiring each \( G(k) \) be strongly connected for all \( k > 0 \). Moreover, to facilitate the analysis of our optimization algorithm, we will make the following assumptions.

**Assumption 2.3:** For each \( i = 1, ..., N \), the function \( f_i \) is differentiable and has Lipschitz continuous gradients, i.e., there exists a constant \( L_i \in (0, +\infty) \) such that

\[
\| \nabla f_i(x) - \nabla f_i(y) \| \leq L_i \| x - y \| \quad \text{for any } x, y \in \mathbb{R}^n
\]

As a consequence, \( f \) is \( \bar{L} \)-smooth with \( \bar{L} = (1/N) \sum_{i=1}^{N} L_i \). We will use \( \hat{L} = \max \{ L_i \} \) in the forthcoming analysis.

**Assumption 2.4:** For each \( i = 1, ..., N \), its objective function \( f_i : \mathbb{R}^n \to \mathbb{R} \) satisfies

\[
f_i(x) \geq f_i(y) + \langle \nabla f_i(y), x - y \rangle + \frac{\mu_i}{2} \| x - y \|^2 \quad \text{for any } x, y \in \mathbb{R}^n
\]

where \( \mu_i \in [0, +\infty) \) and at least one \( \mu_i \) is nonzero. Moreover, we will use \( \hat{\mu} = \max \{ \mu_i \} \) and \( \bar{\mu} = (1/N) \sum_{i=1}^{N} \mu_i \).

**Remark 2.1:** The distributed optimization algorithms we introduced in this paper can be redefined as Push-DIGing algorithm, which exactly incorporates the push-sum protocol [8] into the distributed inexact gradient tracking technique.
III. OPTIMIZATION ALGORITHM

A. The Push-DIGing Algorithm

Based on the above notations and assumptions, we are now ready to propose our main algorithm, namely, Push-DIGing algorithm, to solve the optimization problem (1), followed by its convergence analysis. In our algorithm, each agent \( i \) maintains four variables \( p_i(k), s_i(k), x_i(k) \) and \( y_i(k) \) at each time instant \( k = 1, 2, \ldots \), where \( p_i(k), s_i(k), y_i(k) \) are three auxiliary variables and \( x_i(k) \) is a key variable used to estimate the optimal solution, denoted by \( x^* \). The initialization of our Push-DIGing algorithm chooses an arbitrary \( x_i(0) = p_i(0) \in \mathbb{R}^n, y_i(0) = \nabla f_i(x_i(0)) \in \mathbb{R}, s_i(0) = 1 \) for all \( i = 1, \ldots, N \). Then, every agent \( i \) at each time instant \( k \) updates according to the following rules:

\[
\begin{align*}
p_i(k+1) &= A_i(k)(p_i(k) - \alpha_i y_i(k)) + \sum_{j\in\mathcal{N}_i^0(k)} A_i^j(k)(p_j(k) - \alpha_j y_j(k)) \\
s_i(k+1) &= A_i(k)s_i(k) + \sum_{j\in\mathcal{N}_i^0(k)} A_i^j(k)s_j(k) \\
x_i(k+1) &= p_i(k+1) / s_i(k+1) \\
y_i(k+1) &= A_i(k)(y_i(k) + \nabla f_i(x_i(k+1)) - \nabla f_i(x_i(k))) + \sum_{j\in\mathcal{N}_i^0(k)} A_i^j(k)(y_j(k) + \nabla f_j(x_j(k+1)) - \nabla f_j(x_j(k)))
\end{align*}
\]

(2)

where the positive scalars \( \alpha_i > 0, i = 1, \ldots, N \) are step-sizes, and the vector \( \nabla f_i(x_i(k)) \) is the gradient of the agent \( i \)’s objective function \( f_i(x_i) \) at \( x = x_i(k) \).

Let \( x(t) = [x_1(t), x_2(t), \ldots, x_N(t)]^T \in \mathbb{R}^{nN}, \) and \( \nabla F(x(t)) = [\nabla f_1(x_1(t)), \nabla f_2(x_2(t)), \ldots, \nabla f_N(x_N(t))]^T \in \mathbb{R}^{nN} \). Then the Push-DIGing Algorithm (2) can be rewritten as the following compact matrix-vector form:

\[
\begin{align*}
p(k+1) &= A(k)(p(k) - Dy(k)) \\
s(k+1) &= A(k)s(k), S(k+1) = \text{diag}\{s(k+1)\} \\
x(k+1) &= \left[S(k+1)\right]^{-1} p(k+1) \\
y(k+1) &= A(k)(y(k) + \nabla F(x(k+1)) - \nabla F(x(k)))
\end{align*}
\]

(3)

where \( D \) is a diagonal matrix and \( [D]_{ii} = \alpha_i \) is the constant step-size of agent \( i \).

**Remark 3.1:** Throughout the paper, we only study the case \( n = 1 \) since the analysis method in this paper can be easily extended to multi-dimensional cases.

Here we make a simple algebraic transformation on the Push-DIGing Algorithm (3), i.e.

\[
\begin{align*}
s(k+1) &= A(k)s(k) \\
S(k+1) &= \text{diag}\{s(k+1)\} \\
x(k+1) &= R(k)(x(k) - Dh(k)) \\
h(k+1) &= R(k)h(k) + \left(S(k+1)\right)^{-1} A(k)(\nabla F(x(k+1)) - \nabla F(x(k)))
\end{align*}
\]

(4)
where \( R(k) = (S(k + 1))^{-1} A(k)(S(k)), h(k) = (S(k))^{-1} y(k) \). Noting that, under Assumption 2.1 and 2.2, it is clear that each matrix \( S(k) \) is invertible, and denote \( \| S^{-1} \|_{\text{max}} = \sup_{k \in \mathbb{B}} \| S^{-1}(k) \| \) which can be proven to be bounded. Also, we can prove that \( R(k) \) is actually a row stochastic matrix (see Lemma 4 of [8]).

In what follows, we will introduce the following notation

\[
A_g(k) = A(k)A(k-1)...A(k+1-B)
\]

for any \( k = 0, 1, 2, \ldots \) and \( B = 0, 1, 2, \ldots \) with \( B \leq k + 1 \) and the special case that \( A_g(k) = I \) for any \( k \) and \( A_g(k) = I \) for any needed \( k < 0 \). A crucial property of the norm of \((I - \frac{1}{N}11^T)R_g(k)\) is given in the following lemma, which comes from the properties of push-sum algorithm and can be obtained from references [27, 28].

**Lemma 3.1:** Let Assumptions 2.1 and 2.2 hold, and let \( B \) be an integer satisfying \( B \geq B_g \). Then, for any \( k = B - 1, B, \ldots \) and any matrix \( x \) with appropriate dimensions, if \( x = R_g(k)y \), we have

\[
\| x \| \leq \delta \| y \| \quad \text{where} \quad R_g(k) = (S(k + 1))^{-1} A_g(k)(S(k + 1 - B)) \quad \text{and} \quad \delta = Q_1(1 - \tau^{\frac{1}{N}B_g})^{\frac{2N}{1 - \tau^{\frac{1}{N}B_g}}} < 1, \quad Q_1 = \frac{1}{N^{\frac{1}{2} + \frac{1}{N}b}}.
\]

**Proof:** We omit the proof of Lemma 3.1 since it is almost identical to that of Lemma 13 in [25].

**B. The small gain theorem**

In order to achieve a geometric convergence of the Push-DIGing algorithm, we first introduce a preliminary result, namely, the small gain theorem. It is a somewhat unusual version derived from [25]. Moreover, the original version of the theorem has received a widely research and been extensively employed in control theory [29]. Before stating the small gain theorem, we need to give some notations.

Denote the notation \( u_i \) for the infinite sequence \( u_i = (u_i(0), u_i(1), u_i(2), \ldots) \), where \( u_i(k) \in R^{N_i}, \forall i = 1, 2, \ldots, N \). Furthermore, for any positive integer \( K \), let us define

\[
\| u_i \|^i = \max_{k=0, \ldots, K} \frac{1}{\lambda^k} \| u_i(k) \| \quad \text{and} \quad \| u_i \|^i = \sup_{k \in \mathbb{B}} \frac{1}{\lambda^k} \| u_i(k) \|
\]

where the parameter \( \lambda \in (0, 1) \) will act as the geometric convergence parameter in our later analysis. Based on the above definition, a sufficient condition on the boundedness of \( \| u_i \|^i \) will be given in the small gain theorem that we next stated. As a fundamental result in control systems, more details about the theorem can be found in [29].
Theorem 3.1 (The small gain theorem [25]): Suppose that $u_1, \ldots, u_m$ are sequences such that for all $K > 0$ and each $i = 1, \ldots, m$ we have
\[
\| u_{(i \mod m) + 1} \|^\gamma_i \leq \gamma_i \| u_i \|^\gamma_i + \psi_i
\] (7)
where $\psi_i$ are some constant, and the nonnegative constants (gains) $\gamma_1, \ldots, \gamma_m$ satisfy
\[
\gamma_1 \gamma_2 \cdots \gamma_m < 1
\] (8)

Then, we have
\[
\| u_i \|^\gamma_i \leq \frac{1}{1 - \gamma_2 \cdots \gamma_m} \left( \gamma_1 \gamma_m \cdots \gamma_{i-1} \psi_1 + \gamma_1 \gamma_m \cdots \gamma_{i-2} \psi_2 + \cdots + \gamma_1 \psi_{i-1} + \psi_i \right)
\] (9)

Proof: The proof is omitted here since this is straightforward to derive from [25].

Lemma 3.2: For any matrix sequence $u_i$ and a positive constant $\lambda \in (0,1)$, if $\| u_i \|^\gamma_i$ is bounded, then $\| u_i \|$ converges to 0 with the geometric rate $O(\lambda^k)$.

Proof: Assume that $\| u_i \|^\gamma_i \leq M$, where $M$ is some arbitrary nonnegative constant, then by the definitions we obtain $\sup_{k \geq 0} \frac{1}{\lambda^k} \| u_i(k) \| \leq M$, therefore $\frac{1}{\lambda^k} \| u_i(k) \| \leq M, \forall k$. The proof then follows immediately from $\| u_i(k) \| \leq M \lambda^k, \forall k$.

Before proceeding to the main proof idea, we need to define a few quantities which will use frequently in our analysis. Consider the following additional notations
\[
q(k) = x(k) - 1^\star \quad \text{for any} \quad k = 0, 1, \ldots
\] (10)
\[
z(k) = \nabla F(x(k)) - \nabla F(x(k - 1)) \quad \text{for any} \quad k = 1, 2, \ldots
\] (11)
where $x^\star \in R$ is the optimal solution of problem (1), and the initiation $z(0) = 0$.

Consider the small gain theorem, the geometric convergence of $\| q(k) \|$ can be achieved by applying this theorem to the following circle of arrows:
\[
q \rightarrow z \rightarrow h \rightarrow \left\{ \begin{array}{c}
\tilde{x} \\
y
\end{array} \right\} \rightarrow q
\] (12)

Remark 3.2: We will use the small gain theorem based on the establishment of each arrow. Specially, we need to be aware of the prerequisite that the sequences $\{\| q \|, \| z \|, \| h \|, \| \tilde{x} \|, \| y \|, \| g \| \}$ are proved to be bounded. Thus, we can conclude that all quantities in the above circle of arrows converge at an geometric rate $O(\lambda^k)$. Furthermore, to apply the small gain theorem in the following analysis, we need to require that the product of gains $\gamma_i$ is less than one, which will be achieved by seeking out an appropriate step-size matrix $D$. Now, we are ready to present the establishment of each arrow in the above circle (12).
C. Supporting lemmas

Before introducing the Lemma 3.3, we make some definitions only used in this lemma, which distinguish the notation used in our distributed optimization problem, algorithm and analysis. Problem (1) is redefined as follows with different notation,

\[
\min_{x \in \mathbb{R}^n} g(x) = \frac{1}{N} \sum_{i=1}^{N} g_i(x)
\]  

(13)

where each function \(g_i\) satisfies Assumption 2.1 and 2.2. Consider the following inexact gradient descent on the function \(g\):

\[
v_{k+1} = v_k - \theta \frac{1}{N} \sum_{i=1}^{N} \nabla g_i(u_i(k)) + e_k
\]  

(14)

where \(\theta\) is the step-size and \(e_k\) is an additive noise. Let \(v^*\) be the global optimal solution of \(g\) and define

\[
r_k = \|v_k - v^*\| \quad \text{for any} \ k = 0, 1, ...
\]

Based on the above definitions, we next introduce Lemma 3.3.

**Lemma 3.3:** Suppose that

\[
\beta \geq 2 \quad \text{and} \quad \eta > 0.
\]

Let Assumptions 2.1 and 2.2 hold for every function \(g_i\). For the problem (13), consider the sequences \(\{r_k\}\) and \(\{v_k\}\) be generated by the inexact gradient descent algorithm (14). Then, for any positive integer \(K\), we have

\[
|r_k| \leq 2r_0 + \frac{\sqrt{1 - \frac{\theta \eta \beta}{2(\beta + 1)}}}{\lambda \sqrt{\eta \beta}} \|e\|_K + \frac{1}{\lambda \sqrt{N}} \sqrt{\frac{\hat{L}(1 + \eta) + \frac{\hat{\mu}}{\beta}}{\eta \beta} \sum_{i=1}^{N} \|v - u_i\|_K}
\]  

(15)

**Proof:** We refer the reader to the paper [26] for the proof of Lemma 3.3.

In the following lemma, we start with the first demonstration of the circle (12) which is grounded on the error bound of the inexact gradient descent algorithm in Lemma 3.3.

**Lemma 3.4** (\(\|x\|, \|y\| \rightarrow \|q\|\)): Let Assumptions 2.2, 2.3 and 2.4 hold. Also, assume that the parameters \(\alpha\) and \(\lambda\) satisfy

\[
\sqrt{1 - \frac{\alpha \eta \beta}{2(\beta + 1)}} \leq \lambda < 1 \quad \text{and} \quad 0 < \alpha < \min\left\{\frac{\beta + 1}{\beta \eta \beta}, \frac{1}{\hat{L}(1 + \eta)} \frac{3}{\hat{\mu}}\right\}
\]

where \(\beta \geq 2\) and \(\eta > 0\) are some adjustable parameters. Then, we have that for all \(K = 0, 1, ...\),

\[
|x|_K \leq 2r_0 + \frac{\sqrt{1 - \frac{\alpha \eta \beta}{2(\beta + 1)}}}{\lambda \sqrt{\eta \beta}} \|e\|_K + \frac{1}{\lambda \sqrt{N}} \sqrt{\frac{\hat{L}(1 + \eta) + \frac{\hat{\mu}}{\beta}}{\eta \beta} \sum_{i=1}^{N} \|v - u_i\|_K}
\]  

(15)
\[ \| q \|^{i,K} \leq 2\sqrt{\tilde{N}} \| \pi(0) - x^* \| + (1 + \sqrt{\tilde{N}}) \left( \frac{\tilde{N}}{\lambda} \left( \frac{\tilde{L}(1+\eta) + \tilde{\beta}}{\eta^2} \right) \right) \| \tilde{x} \|^{i,K} + C_{\gamma,a} \| y \|^{i,K} \]  

(16)

where the positive constant \( C_{\gamma,a} = \frac{\sqrt{3} - \alpha_{\max} \beta}{\lambda \tilde{\alpha}} \), \( k = \frac{\alpha_{\max}}{\alpha_{\min}} \) (\( \alpha_{\max} \), \( \alpha_{\min} \) are max, \( \alpha_i \), \( \alpha_{\min} \) min, \( \alpha_i \))

if \( \alpha = \alpha_{\max} \), \( C_{\gamma,a} = \frac{\sqrt{3} - \alpha_{\max} \beta}{\alpha_{\max} \alpha_{\min} \lambda \tilde{\alpha}} \left( \sum_{i=1}^{N} (\alpha_i - \tilde{\alpha})^2 \right) \) if \( \alpha = \tilde{\alpha} \), \( \tilde{\alpha} = \frac{1}{N} \sum_{i=1}^{N} \alpha_i \).

**Proof:** See the Appendix.

Next, to demonstrate the second arrows in the circle (12), we proceed by showing two lemmas. One is \( \| h \| \rightarrow \| \tilde{x} \| \) and the other is \( \| h \| \rightarrow \| y \| \).

**Lemma 3.5** (\( \| h \| \rightarrow \| \tilde{x} \| \)): Let Assumptions 2.1 and 2.2 hold, and let \( \lambda \) be a positive constant in \( (\sqrt{2} \tilde{\alpha}, 1) \), where \( \tilde{\alpha} \) is the constant given in Lemma 3.1. Then, we get

\[ \| \tilde{x} \|^{i,K} \leq \left( \frac{D}{\lambda^2 - \delta} \right) \left( \| \lambda + Q \left( \frac{\lambda - \tilde{\lambda}}{\lambda^2 - \delta} \right) \| \| h \|^{i,K} + \frac{\lambda^2}{\lambda^2 - \delta} \sum_{i=1}^{N} \lambda^{-i-1} \| \tilde{x}(t-1) \| \right) \]  

(17)

for all \( K = 0,1,\ldots \), where \( Q \) is the constant as defined in Lemma 3.1.

**Proof:** Note that \( x(k+1) = R(k)(x(k) - Dh(k)) \). The results can be obtained by the same argument as that illustrated in the proof of Lemma 6 in [25], thus we can achieve (17).

**Lemma 3.6** (\( \| h \| \rightarrow \| y \| \)): For any positive integer \( K \), \( \| y \|^{i,K} \leq \max \| h \|^{i,K} \) holds, where \( \| S \|_{\max} \) is the constant defined above.

**Proof:** Considering \( y(k) = S(k)h(k) \), we have \( \| y(k) \| \leq \| S(k)h(k) \| \leq \| S \|_{\max} \| h(k) \| \). The result immediately follows by applying the corresponding properties of Euclidean norm.

The next lemma presents the establishment of the third arrows in the circle (12).

**Lemma 3.7** (\( \| z \|^{i,K} \rightarrow \| h \|^{i,K} \)): Let Assumptions 2.1-2.4 hold, let the parameter \( \delta \) be as given in Lemma 3.1, and let \( \lambda \) be a positive constant in \( (\sqrt{2} \tilde{\alpha}, 1) \). Then, we have for all \( K = 0,1,\ldots \),

(i) \( \| h \|^{i,K} \leq \tilde{h} \|^{i,K} + \| h \|^{i,K} \)

(ii) \( \| h \|^{i,K} \leq \tilde{Q} \| S^{-1} \|_{\max} \| C \|_{\max} \left( \frac{\lambda(1-\lambda^2)}{(\lambda^2 - \delta)(1-\lambda)} \right) \| \tilde{h} \|^{i,K} + \frac{\lambda^2}{(\lambda^2 - \delta)} \sum_{i=1}^{N} \lambda^{-i-1} \| \tilde{h}(t-1) \| \)

(iii) \( \| h \|^{i,K} \leq \frac{\| JR \|_{\max}}{\lambda} \| h \|^{i,K} + \| S^{-1} \|_{\max} \| C \|_{\max} \| z \|^{i,K} \)

Specially, suppose that \( \| JR \|_{\max} < \lambda < 1 \). Then, followed by above three items, we finally obtain
\[
\| h \|^2 \leq \frac{\| S^{-1} \|_{\max} \| A \|_{\max} (1 + Q) \lambda(1 - \lambda^N)(1 - \lambda)}{1 - \frac{\| JR \|_{\max}}{\lambda}} \left[ \| z \|^2 + \frac{\lambda^N \sum_{i=1}^N \lambda^{-i-1} \| \hat{h}(t-1) \|}{1 - \frac{\| JR \|_{\max}}{\lambda}} \right] (18)
\]

**Proof:** See the Appendix.

The last arrow in the circle (12) demonstrated in the following Lemma is a simple consequence of the fact that the gradient of \( f \) is \( L \)-Lipschitz.

**Lemma 3.8 \((\| q \|^2 \rightarrow \| z \|^2)\):** Let assumption 2.3 holds. Then, we have for all \( K = 0,1,..., \) and any \( 0 < \lambda < 1 \),

\[
\| z \|^2 \leq L(1 + \frac{1}{\lambda}) \| q \|^2
\]

**Proof:** The proof procedure can imitate that of Lemma 5 in [25], and thus it is omitted.

**IV. MAIN RESULTS**

Based on the circle (12) established in the previous section, next we will demonstrate a major result about the geometrical convergence rate estimate for the Push-DIGing algorithm with uncoordinated step-sizes over a time-varying directed graph sequence.

**Theorem 4.1:** Let Assumptions 2.1-2.4, and Lemma 3.1 hold. Let \( \alpha_{\max} = \max \{ \alpha_i \} \) \( (\alpha_{\min} = \min \{ \alpha_i \} ) \) be the largest (smallest) positive entry element of the uncoordinated step-size matrix \( D \) such that

\[
\alpha_{\max} \epsilon \left\{ \left[ \begin{array}{c}
\frac{(1 - \delta)}{2L(1 + \sqrt{N})(1 + 4\sqrt{N}\kappa)(\delta + Q(B - 1))^N} \\
\left(\frac{1}{1 - \frac{\| JR \|_{\max}}{\lambda}}\right)\| S^{-1} \|_{\max} \| A \|_{\max} (BQ_i + (1/2\delta - \delta))
\end{array} \right] \right\}
\]

where \( \kappa = \tilde{L} / \| S \| \) and \( k_D = \alpha_{\max} / \alpha_{\min} \) is the condition number of the step-size matrix \( D \). Suppose that the condition number \( k_D \) is selected such that

\[
k_D < 1 + \frac{\lambda^N \sum_{i=1}^N \lambda^{-i-1} \| \hat{h}(t-1) \|}{4\sqrt{5}\kappa \| S \|_{\max} \| S^{-1} \|_{\max} \| A \|_{\max} (BQ_i + (1/2\delta - \delta)) - (\lambda^N \sum_{i=1}^N \lambda^{-i-1} \| \hat{h}(t-1) \|)}{\lambda^N \sum_{i=1}^N \lambda^{-i-1} \| \hat{h}(t-1) \|}
\]

Then, the sequence \( x(k) \) generated by the Push-DIGing algorithm with uncoordinated step-sizes converges to \( x^* \) at a global geometric rate \( O(\lambda^k) \), where \( \lambda \in (0,1) \) is given by

\[
\lambda = \max \left\{ \sqrt{\delta + \frac{(GF\alpha_{\max} - HK) + \sqrt{(GF\alpha_{\max} - HK)^2 + 4(C + HF)GK\alpha_{\max}}}{2(C + HF)}}, \sqrt{\frac{\alpha_{\max} \| JR \|_{\max}}{3}} \right\}
\]

with
Proof: It is immediately obtained from Lemma 3.4-3.8 that

i) \( \| q \|^{i,K} \leq \gamma_{11} \| \tilde{x} \|^{i,K} + \gamma_{12} \| y \|^{i,K} + \omega_{1} \)

where \( \gamma_{11} = (1+\sqrt{N}) \left( 1+\frac{\sqrt{N}}{\lambda} \left( \frac{\hat{L}(1+\eta)}{\eta \mu} + \frac{\hat{\mu}}{\mu} \right) \right) \) and \( \gamma_{12} = \frac{\sqrt{\beta - \alpha_{\max,\beta}}}{\lambda \mu} \left( 1-k_{D}^{-1} \right) \), \( \omega_{1} = 2\sqrt{N} \| x(0) - \bar{x} \| ; \)

ii) \( \| \tilde{x} \|^{i,K} \leq \gamma_{21} \| h \|^{i,K} + \omega_{21} \)

where \( \gamma_{21} = \frac{\| D \|}{(\lambda^{\beta} - \delta)} \left( \delta + Q_{1} \frac{\lambda - \lambda^{\beta}}{1 - \lambda} \right) \) and \( \omega_{21} = \frac{\lambda^{\beta}}{(\lambda^{\beta} - \delta)} \sum_{i=1}^{a} \lambda^{-i(1-\beta)} \| \bar{x}(t-1) \| ; \)

iii) \( \| y \|^{i,K} \leq \gamma_{22} \| h \|^{i,K} + \omega_{22} \)

where \( \gamma_{22} = \| S \|_{\max} \) and \( \omega_{22} = 0 \);

iv) \( \| h \|^{i,K} \leq \gamma_{3} \| z \|^{i,K} + \omega_{3} \)

where \( \gamma_{3} = \frac{\| S^{-1} \|_{\max} \| A \|_{\max} \left( 1+Q_{1} \frac{\lambda(1-\lambda^{\beta})}{(\lambda^{\beta} - \delta)(1 - \lambda)} \right)}{1 - \| JR \|_{\max}} \) and \( \omega_{3} = \frac{\lambda^{\beta}}{(\lambda^{\beta} - \delta)} \sum_{i=1}^{a} \lambda^{-i(1-\beta)} \| \bar{h}(t-1) \| ; \)

v) \( \| z \|^{i,K} \leq \gamma_{4} \| q \|^{i,K} + \omega_{4} \)

where \( \gamma_{4} = \hat{L}(1+\frac{1}{\lambda}) \) and \( \omega_{4} = 0 \).

Moreover, to use the small gain theorem, we must choose the step-size \( \alpha_{\max} \) such that

\[ (\gamma_{11} \gamma_{21} + \gamma_{12} \gamma_{22}) \gamma_{3} \gamma_{4} < 1 \]  \hspace{1cm} (20)

which means that

\[
\left[ (1+\sqrt{N}) \left( 1+\frac{\sqrt{N}}{\lambda} \left( \frac{\hat{L}(1+\eta)}{\eta \mu} + \frac{\hat{\mu}}{\mu} \right) \right) \frac{\| D \|}{(\lambda^{\beta} - \delta)} \left( \delta + Q_{1} \frac{\lambda - \lambda^{\beta}}{1 - \lambda} \right) + \frac{\sqrt{\beta - \alpha_{\max,\beta}}}{\lambda \mu} \left( 1-k_{D}^{-1} \right) \| S \|_{\max} \right] \\
\times \frac{\| S^{-1} \|_{\max} \| A \|_{\max} \left( 1+Q_{1} \frac{\lambda(1-\lambda^{\beta})}{(\lambda^{\beta} - \delta)(1 - \lambda)} \right)}{1 - \| JR \|_{\max}} \hat{L}(1+\frac{1}{\lambda}) < 1 \]  \hspace{1cm} (21)

where \( \beta \geq 2, \eta > 0 \) and other constraint conditions on parameters that occur in Lemma 3.4, 3.5, and
3.7 are stated as follows:

\[
0 < \alpha_{\max} < \min \left\{ \frac{\beta + 1}{\beta} \frac{1}{L(1 + \eta)} \frac{3}{\mu} \right\} \tag{22}
\]

\[
\sqrt{1 - \frac{\alpha^2 \beta}{2(\beta + 1)}} \leq \lambda < 1 \tag{23}
\]

\[
\sqrt{B} < \lambda < 1 \tag{24}
\]

\[
\| JR \|_{\max} < \lambda < 1 \tag{25}
\]

Noting that \( \beta \geq 2, \eta > 0 \), it follows from (22) that

\[
0 < \alpha_{\max} < \frac{1}{L(1 + \eta)} \tag{26}
\]

Define two specific values for the parameters \( \beta = 2L / \hat{\mu} \) and \( \eta = 1 \) in Lemma 3.4 to obtain some concrete (probably loose) bound on the convergence rate. Furthermore, by using \( 0.5 \leq \lambda < 1 \) and \( (1 - \lambda^\beta) / (1 - \lambda) \leq B \), from relation (21), we obtain

\[
\alpha_{\max} \leq \frac{(\lambda^\beta - \delta)}{2L(1 + \sqrt{N})(1 + 4\sqrt{N}\sqrt{\kappa}(\delta + Q_1(B - 1))) \times \left( \frac{(\lambda^\beta - \delta)(1 - \| JR \|_{\max})}{\| S^{-1} \|_{\max} \| A \|_{\max} (BQ_1 + (\lambda^\beta - \delta)) - 4\sqrt{\kappa}(1 - k_d^{-1}) \| S \|_{\max}} \right) \tag{27}
\]

For convenience, we rewrite formula (27) as

\[
\alpha_{\max} \leq \frac{(\lambda^\beta - \delta)}{G} \left( \frac{C(\lambda^\beta - \delta)}{F(\lambda^\beta - \delta) + K} + H \right) \tag{28}
\]

where \( F, G, C, H, K \) are defined above. In order to guarantee (21) is non-emptiness or the right hand side of (28) is always positive, we require

\[
k_d < 1 + \frac{(\lambda^\beta - \delta)(1 - \| JR \|_{\max})}{4\sqrt{\kappa} \| S \|_{\max} \| S^{-1} \|_{\max} \| A \|_{\max} (BQ_1 + (\lambda^\beta - \delta)) - (\lambda^\beta - \delta)(1 - \| JR \|_{\max})} \tag{29}
\]

Thus, (28) further implies that

\[
\lambda \geq \sqrt\delta + \frac{(GF\alpha_{\max} - HK) + \sqrt{(GF\alpha_{\max} - HK)^2 + 4(C + HF)GK\alpha_{\max}}}{2(C + HF)} \tag{29}
\]

To work out \( \alpha_{\max} \) when \( 0.5 \leq \lambda < 1 \), we need
\[
\alpha_{\text{max}} \leq \frac{(1-\delta)}{2L(1+\sqrt{N})(1+4\sqrt{N}\sqrt{\kappa})(\delta + Q_i(B - 1))} \times \\
\left\{ (1-\delta)(1-LR_{\text{max}}) - 4\sqrt{5N}(1-k_\beta) \parallel S \parallel_{\text{max}} \parallel S^{-1} \parallel_{\text{max}} \parallel A \parallel_{\text{max}} (BQ_i + (1/2^{\alpha} - \delta)) \right. \\
\left. \parallel S^{-1} \parallel_{\text{max}} \parallel A \parallel_{\text{max}} (BQ_i + (1/2^{\alpha} - \delta)) \right\}^{(30)}
\]

Using \( \beta/(\beta + 1) \geq 2/3 \) in (23), it yields
\[
\sqrt{1 - \frac{\alpha_{\text{max}}}{\beta}} \leq \lambda < 1
\] (31)

The desired result of algorithm (2) with uncoordinated step-sizes can be immediately obtained by gathering the multiple conditions for \( \alpha_{\text{max}} \) and \( \lambda \).

**Remark 4.1:** Noting that the choices of \( \beta, \eta, \alpha_{\text{max}}, \) and \( \lambda \) may make the bounds tighter, but here we do not pay attention to this kind of problem. We aim to give an explicit convergence rate \( \lambda \) for the Push-DIGing algorithm.

**Corollary 4.2:** Suppose that all the assumptions and the definitions \( F, G, C, H, K \) stated in Theorem 4.1 hold. Let \( B \) be a large enough integer constant such that
\[
\delta = Q_i(1-\epsilon^{N_k})^{\frac{2-1}{N_k^{1/2}}} < 1, \quad \text{where} \quad Q_i = 2N^{1+\epsilon^{N_k}}, \tau = \frac{1}{N^{2+\epsilon^{N_k}}}
\]
Then, for any maximum step-size \( \alpha_{\text{max}} \in (0, (C + HF)(1-\delta)^2 + HK(1-\delta))/G(F + K) \) (\( \alpha_{\text{max}} = \max\{\alpha_i\} \)), the sequence \( x(k) \) generated by the Push-DIGing algorithm with uncoordinated step-sizes converges to \( 1x^* \) at a global geometric rate \( O(\lambda^k) \), where
\[
\lambda = \sqrt{\left(2\delta(C + HF) - HK\right) + \sqrt{(2\delta(C + HF) - HK)^2 + 4(C + HF)(HK\delta + G(F + K)\alpha_{\text{max}} + (C + HF)\delta^2)}} \\
2(C + HF)
\] if we choose \( \alpha_{\text{max}} \in (0, M] \), and
\[
\lambda = \sqrt{\frac{(2\delta(C + HF) - HK) + \sqrt{(2\delta(C + HF) - HK)^2 + 4(C + HF)(HK\delta + G(F + K)\alpha_{\text{max}} + (C + HF)\delta^2)}}{2(C + HF)}},
\]
where \( M = \frac{3(1-\lambda_{\text{mid}}^2)}{\beta} = \frac{(C + HF)(\lambda_{\text{mid}}^2 - \delta^2) + HK(\lambda_{\text{mid}}^2 - \delta)}{G(F + K)} \) (\( \lambda_{\text{mid}} \) is an intermediate variable detailedly described in the proof).

**Proof:** It directly follows from (28) that
\[
\alpha_{\text{max}} \leq \frac{(C + HF)(\lambda^2 - \delta^2) + HK(\lambda^2 - \delta)}{GF(\lambda^2 - \delta) + GK}
\] (32)
where \( F, G, C, H, K \) are defined in Theorem 4.1. Recall (31) that
\[ \alpha_{\text{max}} \geq \frac{3(1-\lambda^2)}{\mu} \]  \hspace{1cm} (33)

Then, using relations (32) and (33), we can conclude that there exists \( \lambda \in (\sqrt[3]{\delta},1) \) such that

\[ \left\{ \frac{3(1-\lambda^2)}{\mu}, \frac{(C + HF)(\lambda^2 - \delta)^2 + HK(\lambda^2 - \delta)}{GF(\lambda^2 - \delta) + GK} \right\} \neq \emptyset \]  \hspace{1cm} (34)

Here, we study a smaller interval by enlarging the left-side and reducing the right-side of the interval in (34). Since \( B \geq 1 \), we will prove that

\[ \left\{ \frac{3(1-\lambda^2)}{\mu}, \frac{(C + HF)(\lambda^2 - \delta)^2 + HK(\lambda^2 - \delta)}{G(F + K)} \right\} \neq \emptyset \]  \hspace{1cm} (35)

Noting that when \( \lambda \) increases from \( \sqrt[3]{\delta} \) to 1, the left-side of (35) is decreasing from \( \frac{3(1-\delta^2)}{\mu} \) to 0 monotonically while the right-side is increasing from 0 to \( \frac{(C + HF)(1 - \delta)^2 + HK(1 - \delta)}{G(F + K)} \) monotonically. Thus, when \( \lambda \) varies from \( \sqrt[3]{\delta} \) to 1, the critical value \( M \) of the interval in (35) is valid when \( \lambda \) is given by

\[ \lambda = \sqrt{\frac{(\Delta_1(C + HF) - \overline{B}HK) + \sqrt{(\Delta_1(C + HF) - \overline{B}HK)^2 + 8\Delta_1(C + HF)(G(F + K) + \overline{B}HK + 3G(F + K) - \overline{B}(C + HF)\delta^2)}}{2\Delta_1(C + HF) + 3G(F + K)}} \]  \hspace{1cm} (36)

Here, we assume that the value obtained in (36) is \( \lambda_{\text{mid}} \). Thus, if we choose \( \alpha_{\text{max}} \) such that \( \alpha_{\text{max}} \in (0,M) \), we can use \( \lambda = \sqrt{1 - \frac{\alpha_{\text{max}}}{3}} \), while for \( \alpha_{\text{max}} \in \left( M, \frac{(C + HF)(1 - \delta)^2 + HK(1 - \delta)}{G(F + K)} \right) \), we can also set

\[ \lambda = \sqrt{\frac{(\Delta_1(C + HF) - HK) + \sqrt{(\Delta_1(C + HF) - HK)^2 + 8(C + HF)(HK\delta + G(F + K)\alpha_{\text{max}} - (C + HF)\delta^2)}}{2(C + HF)}} \]

The proof is thus completed.

V. NUMERICAL SIMULATION RESULTS

In this section, a numerical example is presented to illustrate the feasibility of the proposed algorithm and the correctness of theoretical analysis throughout the paper. We consider a class of parameter estimation problem in wireless-sensor network. In the simulation, we use five sensors to cooperatively estimate a parameter \( x \). Thus, the optimization problem is to find \( x \in \mathbb{R} \) to minimize

\[ F(x) = \sum_{i=1}^{N} f_i(x) = \sum_{i=1}^{N} \left( a_i + \frac{\|x - c_i\|^2}{b_i} \right), \]

where \( a_i \) is the initial state of local objective \( f_i(x), \) \( b_i \) is the control gain and \( c_i \) is the observation known only to sensor \( i \). Without loss of generality, we
set $a = [1, 2, 3, 4, 5], b = [3.33, 1.67, 1.11, 0.83, 0.67], c = [0.2, 0.4, 0.6, 0.8, 1]$, the uncoordinated step-sizes $D = [0.035, 0.015, 0.025, 0.045, 0.055]$, and the initial condition $x_i(0) \in (0, 1)$ for all $i = 1, ..., 5$.

Then, the simulation results are to be compared with the well-known distributed gradient decent algorithm (DGD) [6], Push-sum algorithm [8] in Fig. 1. It shows that the Push-DIGing algorithm we introduced in this paper has a linear convergence rates while the DGD and Push-sum algorithm only have sublinear convergence rates under the same settings.

![Fig. 1: Time evolution of $\log_{10} \sum_{i=1}^{t} \frac{\| x_i(k) - x^* \|}{\| x_i(0) - x^* \|}$ among Push-DIGing algorithm, Push-sum algorithm and DGD.](image)

**VI. CONCLUSIONS AND FUTURE WORK**

In this paper, the Push-DIGing algorithm for solving the convex optimization problem (1) with uncoordinated step-sizes has been studied in detail. Under some standard assumptions on the objective function and network connectivity, it has been proven that our algorithm is able to achieve a geometrical convergence rate over time-varying directed graphs. We also provided an explicit choice of $\alpha_{\max}$ and derived an explicit convergence rate $\lambda$ based on the small gain theorem. Furthermore, the correctness and effectiveness of our theoretical results are demonstrated by using a numerical example. Future work should include the case of more complex constrained convex optimization problem, and event-triggered communication among agents in the dynamic (time-varying) networks.
REFERENCES

[1] W. Ren, “Consensus based formation control strategies for multi-vehicle systems,” Proceedings of the American Control Conference, 2006: 4237–4242.

[2] M. Rabi, M. Johansson, “A simple peer-to-peer algorithm for distributed optimization in sensor networks,” IEEE Conference on Decision and Control, 2007, 46: 4705-4710.

[3] M. Rabbat, R. Nowak, “Distributed optimization in sensor networks,” Proceedings of 3rd International Symposium on Information Processing in Sensor Networks, 2004: 20–27.

[4] J. Chen, A. Sayed, “Diffusion adaptation strategies for distributed optimization and learning over networks,” IEEE Transactions on Signal Processing, 2012, 60(8): 4289-4305.

[5] S. Zhu, C. Chen, W. Li, B. Yang, X. Guan, “Distributed optimal consensus filter for target tracking in heterogeneous sensor networks,” IEEE transactions on cybernetics, 2013, 43(6): 1963-1976.

[6] A. Nedic, A. Ozdaglar, “Distributed subgradient methods for multi-agent optimization,” IEEE Transactions on Automatic Control, 2009, 54(1): 48–61.

[7] P. Bianchi, J. Jakubowicz, “Convergence of a multi-agent projected stochastic gradient algorithm for non-convex optimization,” IEEE Transactions on Automatic Control, 2013, 58 (2): 391–405.

[8] A. Nedic, A. Ozdaglar, “Distributed optimization over time-varying directed graphs,” IEEE Transactions on Automatic Control, 2015, 60(3): 601-615.

[9] D. Jakovetic, J. Xavier, J. Moura, “Fast distributed gradient methods,” IEEE Transactions on Automatic Control, 2014, 59(5): 1131-1146.

[10] J. Duchi, A. Agarwal, M. Wainwright, “Dual averaging for distributed optimization: Convergence analysis and network scaling,” IEEE Transactions on Automatic Control, 2012, 57(3): 592-606.

[11] I. Matei, J. Baras, “Performance evaluation of the consensus-based distributed subgradient method under random communication topologies,” IEEE Journal of Selected Topics in Signal Processing, 2011, 5(4): 754-771.

[12] S. Ram, A. Nedic, V. Veeravalli, “Distributed stochastic subgradient projection algorithms for convex optimization,” Journal of optimization theory and applications, 2010, 147(3): 516-545.

[13] M. Zhu, S. Martinez, “On distributed convex optimization under inequality and equality constraints,” IEEE Transactions on Automatic Control, 2012, 57(1): 151-164.

[14] I. Lobel, A. Ozdaglar, D. Feijer, “Distributed multi-agent optimization with state-dependent communication,” Mathematical Programming, 2011, 129(2): 255-284.
[15] A. Olshevsky, “Linear time average consensus on fixed graphs and implications for decentralized optimization and multi-agent control,” arXiv preprint arXiv: 2014, 1411.4186.

[16] A. Nedic, “Asynchronous broadcast-based convex optimization over a network,” IEEE Transactions on Automatic Control, 2011, 56(6): 1337-1351.

[17] I. Lobel, A. Ozdaglar, “Convergence analysis of distributed subgradient methods over random networks,” in Proceedings of the 46th IEEE Conference on Communication, Control, and Computing, 2008: 353-360.

[18] P. Yi, Y. Hong, “Quantized subgradient algorithm and data-rate analysis for distributed optimization,” IEEE Transactions on Control of Network Systems, 2014, 1(4): 380–392.

[19] B. Johansson, T. Keviczky, M. Johansson, K. H. Johansson, “Subgradient methods and consensus algorithms for solving convex optimization problems,” in Proceedings of the 47th IEEE Conference on Decision and Control, 2008: 4185-4190.

[20] D. Yuan, S. Xu, H. Zhao, “Distributed primal–dual subgradient method for multiagent optimization via consensus algorithms,” IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2011, 41(6): 1715-1724.

[21] A. Nedic, A. Ozdaglar, P. Parrilo, “Constrained consensus and optimization in multi-agent networks,” IEEE Transactions on Automatic Control, 2010, 55(4): 922-938.

[22] M. Zhu, S. Martinez, “On distributed convex optimization under inequality and equality constraints via primal-dual subgradient methods,” arXiv preprint arXiv, 2010: 1001.2612.

[23] S. S. Ram, A. Nedic, V. V. Veeravalli, “Incremental stochastic subgradient algorithms for convex optimization,” SIAM Journal on Optimization, 2009, 20(2): 691-717.

[24] J. Xu, S. Zhu, Y. Soh, L. Xie, “Augmented distributed gradient methods for multi-Agent optimization under uncoordinated constant stepsizes,” in Proceedings of the 54th IEEE Conference on Decision and Control, 2015: 2055–2060.

[25] A. Nedic, A. Olshevsky, W. Shi, “Achieving geometric convergence for distributed optimization over time-varying graphs,” arXiv preprint arXiv, 2016: 1607.03218.

[26] A. Nedic, A. Olshevsky, W. Shi, C. A. Uribe, “Geometrically convergent distributed optimization with uncoordinated step-sizes,” arXiv preprint arXiv, 2016: 1609.05877.

[27] A. Nedic, A. Ozdaglar, “Cooperative Distributed Multi-agent Optimization,” Convex Optimization in Signal Processing and Communications, 2010, 340.
APPENDIX

Proof of Lemma 3.4: Multiplying $\frac{1}{N^T}$ on the both sides of the last equation of (3) and noting that $A(k)$ is a column stochastic matrix, we then have

$$\overline{y}(k+1) - \frac{1}{N} 1^T \nabla F(x(k+1)) = \overline{y}(k) - \frac{1}{N} 1^T \nabla F(x(k)) \quad (37)$$

where $\overline{y} = \frac{1}{N} 1^T y$. Since $y_i(0) = \nabla f_i(x_i(0)), i \in V$, it follows that

$$\overline{y}(k) = \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x_i(k)) \quad (38)$$

Let us consider the evolution of $\overline{p}(k)$. Multiplying $\frac{1}{N} 1^T$ at the both sides of the first equation of (3), one can obtain

$$\overline{p}(k+1) = \overline{p}(k) - \alpha \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x_i(k)) + (\alpha \frac{1}{N} 1^T - \frac{1}{N} 1^T D) y(k) \quad (39)$$

Applying Lemma 3.3 to the recursion relation of (39), and $0 < \alpha < \frac{3}{\mu}$, we can achieve

$$\| \overline{p} - x^* \|_{i,K} \leq 2 \| \overline{p}(0) - x^* \| + \frac{1}{\lambda \sqrt{N}} \left( \frac{L(1+\eta) + \mu \beta}{\eta \mu} \right) \sum_{i=1}^{N} \| \overline{p} - x_i \|_{i,K}$$

$$\quad + \frac{\sqrt{3-3\alpha\mu}}{N \lambda \mu} \| (1^T - \frac{1^T}{N} D) y \|_{i,K} \quad (40)$$

Let us analyze the summation in the second term of (40). Since $x(k+1) = [S(k+1)]^{-1} s(k+1)$ and $x(0) = p(0)$, we obtain

$$\sum_{i=1}^{N} \| \overline{p} - x_i \|_{i,K} \leq \sum_{i=1}^{N} \| (\overline{p} - \overline{x}) + (\overline{x} - x_i) \|_{i,K} \\
\leq N \| (\overline{p} - \overline{x}) \|_{i,K} + \sum_{i=1}^{N} \| \overline{x} - x_i \|_{i,K}$$

$$\leq N \| (1^T - \frac{1^T}{N} D) x \|_{i,K} + \sum_{i=1}^{N} \| \overline{x} - x_i \|_{i,K}$$

$$= \| (1 - s)^T x \|_{i,K} + \frac{1}{\sqrt{N}} \| \bar{x} \|_{i,K} \quad (41)$$

Since $q(k) = x(k) - 1x^* = x(k) - \bar{x}(k) + 1\bar{x}(k) - 1\overline{p}(k) + 1\overline{p}(k) - 1x^*$, it follows that

$$q(k) = \bar{x}(k) + \frac{1}{N} [l(1 - s(k)) \nabla x(k) + l(\overline{p}(k) - x^*)] \quad (42)$$

Together with (40) and (41), (42) further implies

[28] A. Nedic, A. Ozdaglar, “Convergence rate for consensus with delays,” Journal of Global Optimization, 2010, 47(3): 437-456.

[29] C. Desoer, M. Vidyasagar, “Feedback Systems: Input-Output Properties,” Siam, 2009.
\[ \| q \|_{i,k} \leq \left( 1 + \frac{\sqrt{N}}{\lambda} \sqrt{\frac{\tilde{L}(1+\eta)}{\eta \bar{\mu}}} + \frac{\bar{\mu}}{\lambda \bar{\mu}} \right) \| \tilde{x} \|_{i,k} + 2\sqrt{N} \| \pi(0) - x^* \| + \left( (\sqrt{N})^{-1} + \frac{1}{\lambda} \sqrt{\frac{\tilde{L}(1+\eta)}{\eta \bar{\mu}}} + \frac{\bar{\mu}}{\lambda \bar{\mu}} \right) \| (1-s)^T x \|_{i,k}^2 + \frac{\sqrt{3-a \bar{\mu}}}{\sqrt{N} \lambda \bar{\mu}} \| (I^T - 1^T D) y \|_{i,k}^2. \] 

Finally, let us bound the third term in (43) as follows

\[ \| (1-s(k))^T x(k) \| \leq \| (1-s(k))^T (I - \frac{1}{N} 11^T) x(k) \| \leq \sqrt{N^2 - N} \| \tilde{x}(k) \| \leq N \| \tilde{x}(k) \|. \] 

Substituting (44) into (43) yields the desired results. The proof is thus established.

**Proof of Lemma 3.7:**

(i) \( \| \tilde{h} \|_{i,k} \leq \| \tilde{h} \|_{i,k} + \| h \|_{i,k} \)

Since \( \| h(k) \| \leq \| (I - \frac{1}{N} 11^T + \frac{1}{N} 11^T) h(k) \| \), it thus follows that

\[ \| h(k) \| \leq \| (I - \frac{1}{N} 11^T) h(k) \| + \| \frac{1}{N} 11^T h(k) \| \leq \| h(k) \| \| \| h(k) \|_J. \] 

Multiplying the above relation with \( \lambda^{-k} \), \( k = 0, 1, \ldots \), it yields that

\[ \lambda^{-k} \| h(k) \| \leq \lambda^{-k} \| \tilde{h}(k) \| + \lambda^{-k} \| h(k) \|_J. \]

Taking the maximum over \( k = 0, \ldots, K \) on both side of (46), the desired result follows immediately.

(ii) \( \| \tilde{h} \|_{i,k} \leq Q_1 \| S^{-1} \|_{\max} \| A \|_{\max} \lambda \| (1-\lambda^g) \| \| h \|_{i,k} + \lambda^g \| \sum_{t=1}^g \lambda t \| h(t-1) \| \)

Using \( z(k) = \nabla F(x(k)) - \nabla F(x(k-1)) \), the relation in Push-DIGing algorithm (4) is equivalent to

\[ h(k+1) = R(k) h(k) + (S(k+1))^{-1} A(k) z(k+1) \]

Then, using Lemma 3.1, for all \( k \geq B+1 \), we can achieve that

\[ \| \tilde{h}(k+1) \| = \| \tilde{J} h(k+1) \| \leq \| \tilde{J} R_g(k) h(k+1-B) \| + \| \tilde{J} R_{r_g}(k)(S(k+2-B))^{-1} A(k+1-B) z(k+2-B) \| + \ldots \]

\[ + \| \tilde{J} R_{r_g}(k)(S(k+1))^{-1} A(k) z(k) \| \] 

\[ \leq \delta \| \tilde{J} h(k+1-B) \| + Q_1 \| S^{-1} \|_{\max} \| A(k+1-B) z(k+2-B) \| + \ldots + Q_4 \| S^{-1} \|_{\max} \| A(k-1) z(k) \| \]

\[ + Q_4 \| S^{-1} \|_{\max} \| A(k) z(k+1) \| \]

\[ \leq \delta \| \tilde{J} h(k+1-B) \| + Q_1 \| S^{-1} \|_{\max} \| A \|_{\max} \sum_{t=1}^g \| z(k+2-t) \| \]

where \( Q_1 \) is the constant defined in Lemma 3.1. Multiplying the above relation with \( \lambda^{-k}, k = B+1, B, \ldots \), we have
\[ \lambda^{-k+1} \| \hat{h}(k+1) \| \leq \frac{\delta}{\lambda^2} \lambda^{-(k+1)-b} \| \hat{h}(k+1-B) \| + Q_1 \| S^{-1} \|_{\max} \| A \|_{\max} \sum_{i=1}^{B} \frac{1}{\lambda^{i-1}} \lambda^{-(i+2-k)} \| z(k+2-i) \| \] (49)

In order to use the norm defined in (5), we need to take \( \max_{k=0,1,...,K} \), which in turn requires a relation \( \lambda^{-k+1} \| \hat{h}(k+1) \| \) with \( k = -1,1,...,B-2 \). To get such a relation, without loss of generality, we assume the initial condition of (49) is

\[ \lambda^{-k+1} \| \hat{h}(k+1) \| \leq \lambda^{-k+1} \| \hat{h}(k+1) \|, k = -1,1,...,B \] (50)

Taking the maximum over \( k = -1,1,...,B-2 \) on the both sides of (49) and the maximum over \( k = B-1,...,K \) on the both sides of (50), by combining the obtained two relations, we immediately obtain

\[ \| \tilde{h} \|^{\lambda,K} \leq \frac{\delta}{\lambda^2} \| \tilde{h} \|^{\lambda,K} + Q_1 \| S^{-1} \|_{\max} \| A \|_{\max} \sum_{i=1}^{B} \frac{1}{\lambda^{i-1}} \| z \|^{\lambda,K} + \sum_{i=1}^{B} \lambda^{-(i-1)} \| \tilde{h}(t-1) \| \] (51)

Therefore, we finally get

\[ \| \tilde{h} \|^{\lambda,K} \leq Q_1 \| S^{-1} \|_{\max} \| A \|_{\max} \frac{\lambda(1-\lambda^b)}{(\lambda^b-\delta)(1-\lambda)} \| z \|^{\lambda,K} + \frac{\lambda^b}{(\lambda^b-\delta)} \sum_{i=1}^{B} \lambda^{-(i-1)} \| \tilde{h}(t-1) \| \] (52)

The desired result follows immediately.

(iii) \( \| h \|_d^{\lambda,K} \leq \frac{\| J \|_{\max} \| h \|^{\lambda,K} + \| S^{-1} \|_{\max} \| A \|_{\max} \| z \|^{\lambda,K} }{\lambda} \)

First, multiplying the relation (47) with \( J = \frac{1}{N} 11^T \) and taking Euclidean norm at the both sides of the obtained equality and using the corresponding properties of Euclidean norm, we therefore have

\[ \| Jh(k+1) \| = \| JR(k)h(k) + J(S(k+1))^{-1} A(k)z(k+1) \| \]

\[ \leq \| JR(k)h(k) \| + \| (S(k+1))^{-1} A(k)z(k+1) \| \] (53)

where we have employed \( \| J \| = \frac{1}{N} \| 11^T \| = 1 \) to obtain the last inequality. Multiplying the above relation with \( \lambda^{-k+1}, k = 0,1,..., \), we immediately have

\[ \lambda^{-k+1} \| h(k+1) \| \leq \frac{\| JR \|_{\max} \lambda^{-k} \| h(k) \| + \| S^{-1} \|_{\max} \| A \|_{\max} \lambda^{-k+1} \| z(k+1) \| }{\lambda} \] (54)

Taking \( \max_{k=0,1,...,K-1} \) on the both sides of (54) gives

\[ \| h \|_d^{\lambda,K} \leq \frac{\| JR \|_{\max} \lambda^{-k} \| h \|^{\lambda,K} + \| S^{-1} \|_{\max} \| A \|_{\max} \| z \|^{\lambda,K} }{\lambda} \] (55)

From the conditions (i), (ii) and (iii), we further to show that the rest proof of Lemma 3.7. By combining the preceding items (i), (ii) and (iii), we have for all \( K = 0,1,..., \) and \( \lambda < 1 \).
\[ \| h \|^2, K \leq \frac{\| JR \|_{\text{max}} \| h \|^2, K + \| S^{-1} \|_{\text{max}} \| A \|_{\text{max}} (1 + Q \frac{\lambda(1-\lambda^\beta)}{(\lambda^\beta - \delta)(1 - \lambda)}) \| z \|^2, K}{\lambda} \]

\[ + \frac{\lambda^\beta}{(\lambda^\beta - \delta)} \sum_{i=1}^\beta \lambda^{-(i-1)} \| \tilde{h}(t-1) \| \]

Rearranging the above formula and recalling \( \| JR \|_{\text{max}} < \lambda < 1 \), we finally have

\[ \| h \|^2, K \leq \frac{\| S^{-1} \|_{\text{max}} \| A \|_{\text{max}} (1 + Q \frac{\lambda(1-\lambda^\beta)}{(\lambda^\beta - \delta)(1 - \lambda)}) \| z \|^2, K + \frac{\lambda^\beta}{(\lambda^\beta - \delta)} \sum_{i=1}^\beta \lambda^{-(i-1)} \| \tilde{h}(t-1) \|}{1 - \frac{\| JR \|_{\text{max}}}{\lambda}} \]

This completes the proof.