Preprints of the 21st IFAC World Congress (Virtual)
Berlin, Germany, July 12-17, 2020

Exponential Convergence for Distributed Optimization Under the Restricted Secant Inequality Condition [⋆]

Xinlei Yi [*] Shengjun Zhang [**] Tao Yang [***] Tianyou Chai [***] Karl H. Johansson [*]

[*] *School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, 100 44, Stockholm, Sweden (e-mail: {xinleiy, kallej}@kth.se).*
[**] *Department of Electrical Engineering, University of North Texas, Denton, TX 76203 USA (e-mail: ShengjunZhang@my.unt.edu)*
[***] *State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, 110819, Shenyang, China (e-mail: {yangtao,tychai}@mail.neu.edu.cn)*

Abstract: This paper considers the distributed optimization problem of minimizing a global cost function formed by a sum of local smooth cost functions by using local information exchange. A standard assumption for proving exponential/linear convergence of existing distributed first-order methods is strong convexity of the cost functions. This does not hold for many practical applications. In this paper, we propose a continuous-time distributed primal-dual gradient descent algorithm and show that it converges exponentially to a global minimizer under the assumption that the global cost function satisfies the restricted secant inequality condition. This condition is weaker than strong convexity and the global minimizer is not necessarily unique. Moreover, a discrete-time distributed primal-dual algorithm is developed from the continuous-time algorithm by Euler's approximation method, which also linearly converges to a global minimizer under the same condition. The theoretical results are illustrated by numerical simulations.

Keywords: Distributed optimization, exponential convergence, primal-dual algorithm, restricted secant inequality

1. INTRODUCTION

Distributed optimization has a long history, which can be traced back to Tsitsiklis (1984); Tsitsiklis et al. (1986); Bertsekas and Tsitsiklis (1989). It has gained renewed interests in recent years due to its wide applications in power systems, machine learning, sensor networks, and cyber-physical systems, just to name a few, see Nedić (2015); Yang et al. (2019); Yuan et al. (2019).

When the cost functions are convex, various distributed optimization algorithms have been developed in both discrete- and continuous-time. Most existing discrete-time distributed algorithms are based on consensus and (sub)gradient descent method, see, e.g., Johansson et al. (2008); Nedić and Ozdaglar (2009); Zhu and Martínez (2011); Tsianos et al. (2012); Nedić and Olshevsky (2014); Yang et al. (2017). Distributed (sub)gradient descent algorithms have at most sub-linear convergence rate for diminishing stepsizes. With fixed stepsizes, distributed (sub)gradient descent algorithms converge faster, but only to a neighborhood of an optimal point, see, e.g., Matei and Baras (2011); Yuan et al. (2015). Recent accelerated algorithms with fixed stepsizes use some sort of historical information in the updates.

Continuous-time distributed algorithms can be classified into two classes depending on whether the algorithm uses the first-order gradient information, see, e.g., Wang and Elia (2010); Gharesifard and Cortés (2014); Yu et al. (2016); Kia et al. (2015); Zhang et al. (2017); Li et al. (2018); Yi et al. (2018); Liang et al. (2019) or the second-order Hessian information, see, e.g., Lu and Tang (2012); Wei et al. (2013).

Among these distributed optimization algorithms, a standard assumption for proving exponential (or linear in the language of optimization) convergence is that (local or global) cost functions are strongly convex. For example, Lu and Tang (2012); Yu et al. (2016); Kia et al. (2015); Zhang et al. (2017); Jakovetić et al. (2015); Nedić et al. (2017); Qu and Li (2018, 2019); Xi et al. (2018); Xu et al. (2018); Xin and Khan (2018); Pu et al. (2018); Jakovetić (2019), assumed that each local cost function is strongly convex and Varagnolo et al. (2016); Li et al. (2018); Saadatniaki

[⋆] This work was supported by the Knut and Alice Wallenberg Foundation, the Swedish Foundation for Strategic Research, the Swedish Research Council, and the National Natural Science Foundation of China under grants 61991403, 61991404, and 61991400.

et al. (2018) assumed that the global cost function is strongly convex.

Unfortunately, in many practical applications, such as least squares and logistic regression, the cost functions normally are not strongly convex, see, e.g., Yang et al. (2020). This situation has motivated researchers to consider alternatives to strong convexity. There are some results in centralized optimization. For instance, Necoara et al. (2019) derived linear convergence rates of several centralized first-order methods for solving the smooth convex constrained optimization problem under the quadratic function growth condition and Karimi et al. (2016) established linear convergence rates of centralized proximal-gradient methods for solving the smooth optimization problem under the assumption that the cost function satisfies the Polyak-Łojasiewicz condition. However, to the best of knowledge, there are only few such results in distributed optimization. Shi et al. (2015) proposed the distributed exact first-order algorithm (EXTRA) and established its linear convergence under the conditions that the global cost function is restricted strongly convex and the optimal set is a singleton. Liang et al. (2019) established exponential/linear convergence of the distributed primal-dual gradient descent algorithm for solving smooth convex optimization under the condition that the primal-dual gradient map is metrically subregular which is weaker than the strict and strong convexity.

In this paper, we consider the problem of solving distributed optimization. We first propose a continuous-time distributed primal-dual gradient algorithm and show that it converges exponentially to a global minimizer under the assumption that the global cost function satisfies the restricted secant inequality condition. This condition is weaker than the (restrict) strong convexity condition assumed by Jakovetić et al. (2015); Nedić et al. (2017); Qu and Li (2018, 2019); Xi et al. (2018); Xu et al. (2018); Xin and Khan (2018); Pu et al. (2018); Jakovetić (2019); Varagnolo et al. (2016); Saadatniaki et al. (2018); Zeng and Yin (2017); Xi and Khan (2017); Shi et al. (2015); Lu and Tang (2012); Yu et al. (2016); Kia et al. (2015); Zhang et al. (2017); Li et al. (2018); Yi et al. (2018) since it does not require convexity and the global minimizer is not necessarily unique. This condition is also different from the metric subregularity criterion assumed by Liang et al. (2019). Moreover, we show that the discrete-time counterpart of the proposed continuous-time distributed algorithm, derived from a simple discretization by Euler's method, also converges linearly to a global minimizer under the same condition.

The rest of this paper is organized as follows. Section 2 introduces some preliminaries. Section 3 presents the problem formulation and assumptions. The main results are stated in Sections 4 and 5. Simulations are given in Section 6. Concluding remarks are offered in Section 7.

Notations: $[n]$ denotes the set $\{1,\dots,n\}$ for any positive constant n. $\mathrm{col}(z_1,\dots,z_k)$ is the concatenated column vector of vectors $z_i \in \mathbb{R}^{p_i}$, $i \in [k]$. $\mathbf{1}_n$ ($\mathbf{0}_n$) denotes the column one (zero) vector of dimension n. $\mathbf{I}_n$ is the n-dimensional identity matrix. Given a vector $[x_1,\dots,x_n]^\top \in \mathbb{R}^n$, $\mathrm{diag}([x_1,\dots,x_n])$ is a diagonal matrix with the i-th diagonal element being x_i. The notation $A\otimes B$ denotes the Kronecker product of matrices A and B. $\mathrm{null}(A)$ is the null space of matrix A. Given two symmetric matrices M, N, $M \geqslant N$ means that $M-N$ is positive semi-definite. $\rho(\cdot)$ stands for the spectral radius for matrices and $\rho_2(\cdot)$ indicates the minimum positive eigenvalue for matrices having positive eigenvalues. $\|\cdot\|$ represents the Euclidean norm for vectors or the induced 2-norm for matrices. For given positive semi-definite matrix A, $\|x\|_A$ denotes the norm $\sqrt{x^\top Ax}$. Given a differentiable function f, ∇f denotes the gradient of f.

2. PRELIMINARIES

In this section, we present some definitions from algebraic graph theory (see Mesbahi and Egerstedt (2010)), the restricted secant inequality (see Zhang and Cheng (2015)), and monotonicity properties of vector functions (see Crouzeix et al. (2000)).

2.1 Algebraic Graph Theory

Let $\mathcal{G} = (\mathcal{V},\mathcal{E},A)$ denote a weighted undirected graph with the set of vertices (nodes) $\mathcal{V} = [n]$, the set of links (edges) $\mathcal{E} \subseteq \mathcal{V}\times\mathcal{V}$, and the weighted adjacency matrix $A = A^\top = (a_{ij})$ with nonnegative elements a_{ij}. A link of $\mathcal{G}$ is denoted by $(i,j) \in \mathcal{E}$ if $a_{ij} > 0$, i.e., if vertices i and j can communicate with each other. It is assumed that $a_{ii} = 0$ for all $i \in [n]$. Let $\mathcal{N}_i = \{j \in [n] : \ a_{ij} > 0\}$ and $\mathrm{deg}_i = \sum_{j=1}^{n} a_{ij}$ denotes the neighbor set and weighted degree of vertex i, respectively. The degree matrix of graph $\mathcal{G}$ is $\mathrm{Deg} = \mathrm{diag}([\mathrm{deg}_1,\cdots,\mathrm{deg}_n])$. The Laplacian matrix is $L = (L_{ij}) = \mathrm{Deg} - A$. A path of length k between vertices i and j is a subgraph with distinct vertices $i_0 = i,\dots,i_k = j \in [n]$ and edges $(i_j, i_{j+1}) \in \mathcal{E}$, $j = 0,\dots,k-1$. An undirected graph is connected if there exists at least one path between any two vertices.

2.2 Restricted Secant Inequality

Definition 1. (Definitions 1 and 2 in Zhang and Cheng (2015)) A differentiable function $f(x) : \mathbb{R}^p \mapsto \mathbb{R}$ satisfies the restricted secant inequality condition with constant $\nu > 0$ if

$$(\nabla f(x) - \nabla f(\mathcal{P}_{X^*}(x)))^\top (x - \mathcal{P}_{X^*}(x)) \geqslant \nu\|x - \mathcal{P}_{X^*}(x)\|^2, \ \forall x \in \mathbb{R}^p, \tag{1}$$

where X^* is the set of all global minimizers of f and $\mathcal{P}_{X^*}(x)$ is the projection of x onto the set X^*, i.e., $\mathcal{P}_{X^*}(x) = \arg\min_{y\in X^*}\|x-y\|^2$. If the function f is also convex it is called restricted strong convexity.

Note that, unlike the strong convexity, the restricted secant inequality (1) alone does not even imply the convexity of f. Moreover, it does not imply that X^* is a singleton either. However, it implies that every stationary point is a global minimizer, i.e., $X^* = \{x \in \mathbb{R}^p : \nabla f(x) = \mathbf{0}_p\}$. Therefore, it is weaker than the (essential and weak) strong convexity, see Karimi et al. (2016).

2.3 Monotonicity

Definition 2. (See Section 2.2 in Crouzeix et al. (2000)) A mapping $F : \mathbb{K} \subseteq \mathbb{R}^p \to \mathbb{R}^p$ is said to be

(1) pseudomonotone on $\mathbb{K}$ if for all $a, b \in \mathbb{K}$,
$$(a-b)^\top F(b) \geqslant 0 \Rightarrow (a-b)^\top F(a) \geqslant 0;$$
(2) pseudomonotone$^+_*$ on $\mathbb{K}$ if it is pseudomonotone on $\mathbb{K}$ and for all $a, b \in \mathbb{K}$,
$$[(a-b)^\top F(b) = 0 \text{ and } (a-b)^\top F(a) = 0] \Rightarrow F(a) = F(b).$$

The gradient of a differentiable pseudoconvex function is pseudomonotone, see Karamardian (1976); Penot and Quang (1997), and the gradient of a differentiable G-convex function is pseudomonotone$^+_*$, see Crouzeix et al. (2000).

3. PROBLEM FORMULATION AND ASSUMPTIONS

Consider a network of n agents, each of which has a local cost function $f_i : \mathbb{R}^p \to \mathbb{R}$. All agents collaborate together to find an optimizer x^* that minimizes the global objective $f(x) = \sum_{i=1}^n f_i(x)$, i.e.,
$$\min_{x \in \mathbb{R}^p} f(x). \tag{2}$$
The communication among agents is described by a weighted undirected graph $\mathcal{G}$. Let X^* denote the optimal set of the optimization problem (2). For simplicity, let $\boldsymbol{x} = \mathrm{col}(x_1, \dots, x_n)$, $\tilde{f}(\boldsymbol{x}) = \sum_{i=1}^n f_i(x_i)$, $\boldsymbol{X}^* = \{\mathbf{1}_n \otimes x^* : x^* \in X^*\}$, and $\boldsymbol{L} = L \otimes \mathbf{I}_p$. The following assumptions are made.

Assumption 1. Each local cost function is differentiable. Moreover, the optimal set X^* is nonempty and convex.

Assumption 2. Each local cost function is smooth, that is, for each $i \in [n]$, f_i has a globally Lipschitz-continuous gradient with constant $L_{f_i} > 0$:
$$\|\nabla f_i(a) - \nabla f_i(b)\| \leqslant L_{f_i} \|a - b\|, \ \forall a, b \in \mathbb{R}^p.$$

Assumption 3. The global cost function $f(x)$ satisfies the restricted secant inequality condition with constant $\nu > 0$.

Assumption 4. $\{\nabla \tilde{f}(\boldsymbol{x}) : \ \boldsymbol{x} \in \boldsymbol{X}^*\}$ is a singleton.

Remark 1. Assumptions 1–2 are mild since the convexity of the cost functions and the boundedness of their gradients are not assumed. Assumption 3 only requires that the global cost function satisfies the restricted secant inequality condition, so it is weaker than the assumptions that the global or each local cost function is strongly convex, which are commonly assumed in the literature. One sufficient condition which satisfies Assumption 4 is that X^* is a singleton. The following lemma gives another sufficient condition. Both sufficient conditions do not require the cost functions to be convex.

Lemma 1. (Proposition 14 in Crouzeix et al. (2000)) Let $\mathbb{H} = \{\mathbf{1}_n \otimes x : \ x \in \mathbb{R}^p\}$. Suppose that each local cost function is differentiable and X^* is nonempty. If $\nabla \tilde{f}$ is pseudomonotone$^+_*$ on $\mathbb{H}$, then $\{\nabla \tilde{f}(\boldsymbol{x}) : \ \boldsymbol{x} \in \boldsymbol{X}^*\}$ is a singleton.

To end this section, we make the following standard assumption on the underlying communication graph.

Assumption 5. The undirected graph $\mathcal{G}$ is connected.

4. CONTINUOUS-TIME DISTRIBUTED ALGORITHM

In this section, we propose a continuous-time distributed algorithm and analyses its convergence rate. Due to the space limitations, all proofs are omitted, but can be found in Yi et al. (2019).

Noting that the Laplacian matrix L is positive semi-definite and $\mathrm{null}(L) = \{\mathbf{1}_n\}$ since $\mathcal{G}$ is connected, we know that the optimization problem (2) is equivalent to the following constrained optimization problem
$$\begin{aligned} \min_{\boldsymbol{x} \in \mathbb{R}^{np}} \quad & \tilde{f}(\boldsymbol{x}) \\ \text{s.t.} \quad & \boldsymbol{L}^{1/2}\boldsymbol{x} = \mathbf{0}_{np}. \end{aligned} \tag{3}$$

Let $\boldsymbol{u} = \mathrm{col}(u_1, \dots, u_n) \in \mathbb{R}^{np}$ denote the dual variable, then the augmented Lagrangian function associated with (3) is
$$\mathcal{A}(\boldsymbol{x}, \boldsymbol{u}) = \tilde{f}(\boldsymbol{x}) + \frac{\alpha}{2}\boldsymbol{x}^\top \boldsymbol{L}\boldsymbol{x} + \beta \boldsymbol{u}^\top \boldsymbol{L}^{1/2}\boldsymbol{x}, \tag{4}$$
where $\alpha > 0$ and $\beta > 0$ are constants. Although $\tilde{f}(\boldsymbol{x})$ may not satisfy the restricted secant inequality condition, the following lemma shows that $\tilde{f}(\boldsymbol{x}) + \frac{\alpha}{2}\boldsymbol{x}^\top \boldsymbol{L}\boldsymbol{x}$ satisfies the restricted secant inequality condition with respect to $\boldsymbol{X}^*$.

Lemma 2. Suppose that Assumptions 1–3 and 5 hold. If $\alpha > \frac{2nL_f^2 + \nu L_f}{\nu \rho_2(L)}$, where $L_f = \max_{i \in [n]}\{L_{f_i}\}$, then
$$\begin{aligned} &(\nabla \tilde{f}(\boldsymbol{x}) - \nabla \tilde{f}(\mathcal{P}_{\boldsymbol{X}^*}(\boldsymbol{x})))^\top (\boldsymbol{x} - \mathcal{P}_{\boldsymbol{X}^*}(\boldsymbol{x})) + \alpha \|\boldsymbol{x}\|^2_{\boldsymbol{L}} \\ &\geqslant \nu_1 \|\boldsymbol{x} - \mathcal{P}_{\boldsymbol{X}^*}(\boldsymbol{x})\|^2, \ \forall \boldsymbol{x} \in \mathbb{R}^{np}, \end{aligned} \tag{5}$$
where $\nu_1 = \min\{\frac{\nu}{2n}, \alpha\rho_2(L) - \frac{2nL_f^2 + \nu L_f}{\nu}\} > 0$.

Remark 2. Lemma 2 extends Proposition 3.6 in Shi et al. (2015) and plays an important role in the proof of the exponential convergence later. The key difference between Lemma 2 and Proposition 3.6 in Shi et al. (2015) is that here we do not assume that $\tilde{f}$ is convex and X^* is a singleton. The requirement that $\alpha > \frac{2nL_f^2 + \nu L_f}{\nu \rho_2(L)}$ is used to eliminate the effects of non-convexity of $\tilde{f}$. Similar to the proof of Proposition 3.6 in Shi et al. (2015), we can show that if $\tilde{f}$ is convex, then this requirement can be relaxed by $\alpha > 0$ and (5) still holds with $\nu_1 = \min\{\frac{\nu}{n} - 2L_f\iota, \frac{\alpha\rho_2(L)\iota^2}{1+\iota^2}\} > 0$, where $\iota \in (0, \frac{\nu}{2nL_f})$. Due to the similarity, we omit the details here.

Based on the primal-dual gradient method, a continuous-time distributed algorithm to solve (3) is
$$\dot{\boldsymbol{x}}(t) = -\alpha \boldsymbol{L}\boldsymbol{x}(t) - \beta \boldsymbol{L}^{1/2}\boldsymbol{u}(t) - \nabla \tilde{f}(\boldsymbol{x}(t)), \tag{6a}$$
$$\dot{\boldsymbol{u}}(t) = \beta \boldsymbol{L}^{1/2}\boldsymbol{x}(t), \ \forall \boldsymbol{x}(0), \ \boldsymbol{u}(0) \in \mathbb{R}^{np}. \tag{6b}$$

Denote $\boldsymbol{v} = \mathrm{col}(v_1, \dots, v_n) = \boldsymbol{L}^{1/2}\boldsymbol{u}$, then the algorithm (6) can be rewritten as
$$\dot{\boldsymbol{x}}(t) = -\alpha \boldsymbol{L}\boldsymbol{x}(t) - \beta \boldsymbol{v}(t) - \nabla \tilde{f}(\boldsymbol{x}(t)), \tag{7a}$$
$$\dot{\boldsymbol{v}}(t) = \beta \boldsymbol{L}\boldsymbol{x}(t), \ \forall \boldsymbol{x}(0) \in \mathbb{R}^{np}, \ \boldsymbol{v}(0) = \mathbf{0}_{np}, \tag{7b}$$
or
$$\dot{x}_i(t) = -\alpha \sum_{j=1}^n L_{ij} x_j(t) - \beta v_i(t) - \nabla f_i(x_i(t)), \tag{8a}$$
$$\dot{v}_i(t) = \beta \sum_{j=1}^n L_{ij} x_j(t), \ \forall x_i(0) \in \mathbb{R}^p, \ v_i(0) = \mathbf{0}_p. \tag{8b}$$

We have the following result for the continuous-time distributed primal-dual gradient descent algorithm (8).

Theorem 1. Each agent $i \in [n]$ runs the distributed algorithm (8). If Assumptions 1–5 hold, $\alpha > \frac{2nL_f^2+\nu L_f}{\nu\rho_2(L)}$, and $\beta > 0$, then $\|\boldsymbol{x}(t) - \mathcal{P}_{\boldsymbol{X}^*}(\boldsymbol{x}(t))\|$ exponentially converges to 0 with a rate no less than $\frac{\epsilon_2}{2\epsilon_3} > 0$, where $\epsilon_2 = \min\{\frac{\beta}{2}, \epsilon_1\nu_1\} > 0$ and $\epsilon_3 = \max\{\frac{\epsilon_1}{\rho_2(L)} + \frac{\alpha}{2\beta} + \frac{1}{2}, \epsilon_1 + \frac{1}{2}\}$, with $\epsilon_1 = \max\{\frac{1}{\nu_1}(\frac{L_f^2}{2\beta} + \rho(L)\beta), \frac{\beta}{\alpha}\}$.

Remark 3. The exponential convergence for continuous-time distributed algorithms was also established by Lu and Tang (2012); Yu et al. (2016); Kia et al. (2015); Zhang et al. (2017); Li et al. (2018); Yi et al. (2018); Liang et al. (2019). However, Lu and Tang (2012); Yu et al. (2016); Kia et al. (2015); Zhang et al. (2017) assumed that each local cost function is strongly convex. Li et al. (2018) assumed that the global cost function is strongly convex. Yi et al. (2018) assumed that the global cost function is restricted strongly convex and the optimal set is a singleton. Liang et al. (2019) assumed that each local cost function is convex and the primal-dual gradient map is metrically subregular. In contrast, the exponential convergence result established in Theorem 1 only requires that the global cost function satisfies the restricted secant inequality condition, but the convexity assumption on cost functions and the singleton assumption on the optimal set are not required.

5. DISCRETE-TIME DISTRIBUTED ALGORITHM

In this section, we propose a discrete-time distributed algorithm and analyse its convergence rate.

Consider a discretization of the continuous-time algorithm (7) by Euler's approximation method as

$$\boldsymbol{x}(k+1) = \boldsymbol{x}(k) - h(\alpha \boldsymbol{L}\boldsymbol{x}(k) + \beta\boldsymbol{v}(k) + \nabla\tilde{f}(\boldsymbol{x}(k))), \quad (9a)$$

$$\boldsymbol{v}(k+1) = \boldsymbol{v}(k) + h\beta\boldsymbol{L}\boldsymbol{x}(k), \ \forall\boldsymbol{x}(0) \in \mathbb{R}^{np}, \ \boldsymbol{v}(0) = \boldsymbol{0}_{np}, \quad (9b)$$

where $h > 0$ is a fixed stepsize. It is straightforward to check that the algorithm (9) is equivalent to the algorithm EXTRA proposed in Shi et al. (2015) with mixing matrices $W = \mathbf{I}_{np} - h\alpha\boldsymbol{L}$ and $\tilde{W} = \mathbf{I}_{np} - h\alpha\boldsymbol{L} + h^2\beta^2\boldsymbol{L}$. The distributed form of (9) is

$$x_i(k+1) = x_i(k) - h(\alpha\sum_{j=1}^{n} L_{ij}x_j(k) + \beta v_i(k) + \nabla f_i(x_i(k))), \quad (10a)$$

$$v_i(k+1) = v_i(k) + h\beta\sum_{j=1}^{n} L_{ij}x_j(k), \quad \forall x_i(0) \in \mathbb{R}^p, \ v_i(0) = \boldsymbol{0}_p. \quad (10b)$$

We have the following result for the discrete-time distributed primal-dual gradient descent algorithm (10).

Theorem 2. Each agent $i \in [n]$ runs the distributed algorithm (10). If Assumptions 1–5 hold, $\alpha > \frac{2nL_f^2+\nu L_f}{\nu\rho_2(L)}$, $\beta > 0$, and $0 < h < \frac{2\epsilon_2\epsilon_4}{\eta\epsilon_3\epsilon_5}$, where $\eta = \sqrt{2}\max\{\frac{2\epsilon_1}{\rho_2(L)} + \alpha + 1, 4\epsilon_1 + 1\} > 0$, $\epsilon_4 = \epsilon_1\min\{\frac{1}{\rho(L)}, \frac{1}{2}\}$, and $\epsilon_5 = \max\{\beta^2\rho^2(L) + 3\alpha^2\rho^2(L) + 3L_f^2, 3\beta^2\}$, then $\|\boldsymbol{x}(k) - \mathcal{P}_{\boldsymbol{X}^*}(\boldsymbol{x}(k))\|$ linearly converges to 0 with a rate no less than $1 - \gamma$, where $\gamma = \frac{h(2\epsilon_2\epsilon_4 - h\eta\epsilon_3\epsilon_5)}{4\epsilon_3\epsilon_4}$.

Remark 4. By comparing Theorems 1 and 2, we see that the proposed continuous- and discrete-time distributed algorithms have the same convergence properties under the same assumptions. The linear convergence for discrete-time distributed algorithms was also established by Jakovetić et al. (2015); Nedić et al. (2017); Qu and Li (2018, 2019); Xi et al. (2018); Xu et al. (2018); Xin and Khan (2018); Pu et al. (2018); Jakovetić (2019); Varagnolo et al. (2016); Saadatniaki et al. (2018); Zeng and Yin (2017); Xi and Khan (2017); Shi et al. (2015). However, Jakovetić et al. (2015); Nedić et al. (2017); Qu and Li (2018, 2019); Xi et al. (2018); Xu et al. (2018); Xin and Khan (2018); Pu et al. (2018); Jakovetić (2019) assumed that each local cost function is strongly convex; Varagnolo et al. (2016); Saadatniaki et al. (2018) assumed that the global cost function is strongly convex. Zeng and Yin (2017); Xi and Khan (2017) assumed that each local cost function is restricted strongly convex and the optimal set X^* is a singleton. Shi et al. (2015) assumed that the global cost function is restricted strongly convex and X^* is a singleton. In contrast, the linear convergence result established in Theorem 2 only requires that the global cost function satisfies the restricted secant inequality condition, but the convexity assumption on cost functions and the singleton assumption on the optimal set are not required.

6. SIMULATIONS

In this section, we verify the theoretical results through a numerical example. Consider the distributed optimization problem (2) with

$$f_i(x) = \begin{cases} b_{i,1}(x+1)^2, & x \leqslant -1, \\ b_{i,2}x^4, & -1 < x \leqslant 0, \\ 1 - \sqrt{1-x^2} + b_{i,3}x^2, & 0 \leqslant x < \frac{\sqrt{2}}{2}, \\ f_{i,1}(x), & \frac{\sqrt{2}}{2} \leqslant x < 1, \\ f_{i,2}(x), & x \geqslant 1, \end{cases}$$

where $f_{i,1}(x) = \sqrt{1-(x-\sqrt{2})^2} - \sqrt{2} + 1 + b_{i,3}x^2$, $f_{i,2}(x) = \frac{1}{2}(x - 1 + \sqrt{\frac{\sqrt{2}-1}{2}})^2 + \sqrt{2\sqrt{2}-2} + \frac{5-5\sqrt{2}}{4} + b_{i,3}x^2$, and $b_{i,j}$, $j = 1,2,3$ are constants that are randomly generated and satisfy the condition that $\sum_{i=1}^n b_{i,1} > 0$ and $\sum_{i=1}^n b_{i,2} = \sum_{i=1}^n b_{i,3} = 0$. These $f_i(x)$, $i \in [n]$ are modifications of Example 2 in Zhang and Cheng (2015). Clearly, f_i is non-convex but differentiable and smooth, and the global objective $f(x) = \sum_{i=1}^n f_i(x)$ satisfies the restricted secant inequality condition with constant $\nu = \min\{\sqrt{\frac{\sqrt{2}-1}{2}}, 2\sum_{i=1}^n b_{i,1}\}$, see Zhang and Cheng (2015). Moreover, the optimal set is $[-1, 0]$. The communication graph between agents is modeled as a ring graph with $n = 10$ agents.

We run the discrete-time distributed algorithm (10) with $\alpha = \beta = 10$ and $h = 0.02$. The initial value $x_i(0)$ is randomly generated. The trajectories of the primal and dual variables of each agent are plotted in Fig. 1 and Fig. 2, respectively. We see that each primal variable converges to zero which is a global minimizer and correspondingly each dual variable also converges to zero. Evolutions of residual $\|\boldsymbol{x}(k) - \mathcal{P}_{\boldsymbol{X}^*}(\boldsymbol{x}(k))\|/\|\boldsymbol{x}(0) - \mathcal{P}_{\boldsymbol{X}^*}(\boldsymbol{x}(0))\|$ are shown in Fig. 3. The results illustrate linear convergence, which are consistent with the theoretical results of Theorem 2.

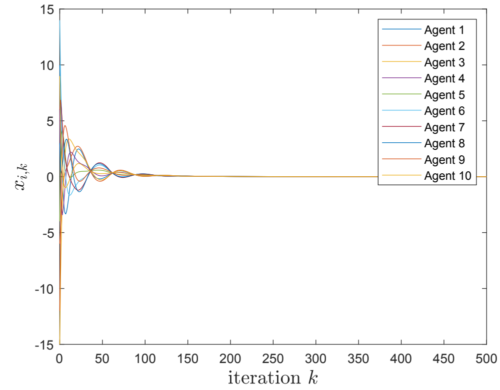

Fig. 1. Evolutions of local primal variables.

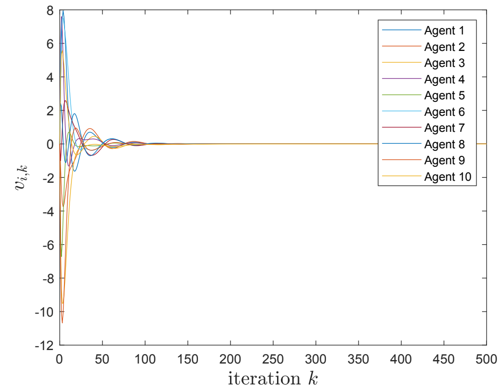

Fig. 2. Evolutions of local dual variables.

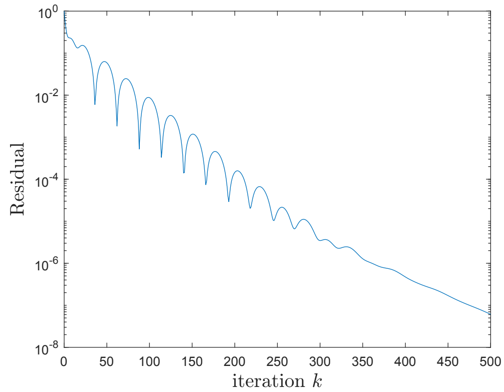

Fig. 3. Evolutions of residual.

7. CONCLUSIONS

In this paper, we derived the exponential convergence rate of the continuous-time distributed primal-dual algorithm for solving the distributed smooth optimization problem when the global cost function satisfies the restricted secant inequality condition. This condition relaxes the standard strong convexity condition. We also showed that the discrete-time counterpart of the continuous-time algorithm establishes linear convergence rate under the same condition. An interesting future research direction is to relax the restricted secant inequality condition by the Polyak-Łojasiewicz condition.

REFERENCES

Bertsekas, D.P. and Tsitsiklis, J.N. (1989). *Parallel and Distributed Computation: Numerical Methods*. Englewood Cliffs, NJ: Prentice Hall.

Crouzeix, J.P., Marcotte, P., and Zhu, D. (2000). Conditions ensuring the applicability of cutting-plane methods for solving variational inequalities. *Mathematical Programming*, 88(3), 521–539.

Gharesifard, B. and Cortés, J. (2014). Distributed continuous-time convex optimization on weight-balanced digraphs. *IEEE Transactions on Automatic Control*, 59(3), 781–786.

Jakovetić, D. (2019). A unification and generalization of exact distributed first-order methods. *IEEE Transactions on Signal and Information Processing over Networks*, 5(1), 31–46.

Jakovetić, D., Moura, J.M., and Xavier, J. (2015). Linear convergence rate of a class of distributed augmented Lagrangian algorithms. *IEEE Transactions on Automatic Control*, 60(4), 922–936.

Johansson, B., Keviczky, T., Johansson, M., and Johansson, K.H. (2008). Subgradient methods and consensus algorithms for solving convex optimization problems. In *IEEE Conference on Decision and Control*, 4185–4190.

Karamardian, S. (1976). Complementarity problems over cones with monotone and pseudomonotone maps. *Journal of Optimization Theory and Applications*, 18(4), 445–454.

Karimi, H., Nutini, J., and Schmidt, M. (2016). Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 795–811.

Kia, S.S., Cortés, J., and Martínez, S. (2015). Distributed convex optimization via continuous-time coordination algorithms with discrete-time communication. *Automatica*, 55, 254–264.

Li, Z., Ding, Z., Sun, J., and Li, Z. (2018). Distributed adaptive convex optimization on directed graphs via continuous-time algorithms. *IEEE Transactions on Automatic Control*, 63(5), 1434–1441.

Liang, S., Wang, L.Y., and Yin, G. (2019). Exponential convergence of distributed primal–dual convex optimization algorithm without strong convexity. *Automatica*, 105, 298–306.

Lu, J. and Tang, C.Y. (2012). Zero-gradient-sum algorithms for distributed convex optimization: The continuous-time case. *IEEE Transactions on Automatic Control*, 57(9), 2348–2354.

Matei, I. and Baras, J.S. (2011). Performance evaluation of the consensus-based distributed subgradient method under random communication topologies. *IEEE Journal of Selected Topics in Signal Processing*, 5(4), 754–771.

Mesbahi, M. and Egerstedt, M. (2010). *Graph Theoretic Methods in Multiagent Networks*. Princeton University Press.

Necoara, I., Nesterov, Y., and Glineur, F. (2019). Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175(1-2), 69–107.

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

Nedić, A. (2015). Convergence rate of distributed averaging dynamics and optimization in networks. *Foundations and Trends in Systems and Control*, 2(1), 1–100.

Nedić, A. and Olshevsky, A. (2014). Distributed optimization over time-varying directed graphs. *IEEE Transactions on Automatic Control*, 60(3), 601–615.

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

Penot, J.P. and Quang, P. (1997). Generalized convexity of functions and generalized monotonicity of set-valued maps. *Journal of Optimization Theory and Applications*, 92(2), 343–356.

Pu, S., Shi, W., Xu, J., and Nedić, A. (2018). A push-pull gradient method for distributed optimization in networks. In *IEEE Conference on Decision and Control*, 3385–3390.

Qu, G. and Li, N. (2018). Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3), 1245–1260.

Qu, G. and Li, N. (2019). Accelerated distributed Nesterov gradient descent. *IEEE Transactions on Automatic Control*. doi:10.1109/TAC.2019.2937496.

Saadatniaki, F., Xin, R., and Khan, U.A. (2018). Optimization over time-varying directed graphs with row and column-stochastic matrices. *arXiv preprint arXiv:1810.07393*.

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

Tsianos, K.I., Lawlor, S., and Rabbat, M.G. (2012). Push-sum distributed dual averaging for convex optimization. In *IEEE Conference on Decision and Control*, 5453–5458.

Tsitsiklis, J.N. (1984). *Problems in decentralized decision making and computation*. Ph.D. thesis, MIT, Cambridge, MA.

Tsitsiklis, J., Bertsekas, D., and Athans, M. (1986). Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Transactions on Automatic Control*, 31(9), 803–812.

Varagnolo, D., Zanella, F., Cenedese, A., Pillonetto, G., and Schenato, L. (2016). Newton-Raphson consensus for distributed convex optimization. *IEEE Transactions on Automatic Control*, 61(4), 994–1009.

Wang, J. and Elia, N. (2010). Control approach to distributed optimization. In *the Annual Allerton Conference on Communication, Control, and Computing*, 557–561.

Wei, E., Ozdaglar, A., and Jadbabaie, A. (2013). A distributed Newton method for network utility maximization–I: Algorithm. *IEEE Transactions on Automatic Control*, 58(9), 2162–2175.

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

Xi, C., Xin, R., and Khan, U.A. (2018). ADD-OPT: Accelerated distributed directed optimization. *IEEE Transactions on Automatic Control*, 63(5), 1329–1339.

Xin, R. and Khan, U.A. (2018). A linear algorithm for optimization over directed graphs with geometric convergence. *IEEE Control Systems Letters*, 2(3), 325–330.

Xu, J., Zhu, S., Soh, Y.C., and Xie, L. (2018). Convergence of asynchronous distributed gradient methods over stochastic networks. *IEEE Transactions on Automatic Control*, 63(2), 434–448.

Yang, T., George, J., Qin, J., Yi, X., and Wu, J. (2020). Distributed least squares solver for network linear equations. *Automatica*, 113, 108798.

Yang, T., Lu, J., Wu, D., Wu, J., Shi, G., Meng, Z., and Johansson, K.H. (2017). A distributed algorithm for economic dispatch over time-varying directed networks with delays. *IEEE Transactions on Industrial Electronics*, 64(6), 5095–5106.

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

Yi, X., Yao, L., Yang, T., George, J., and Johansson, K.H. (2018). Distributed optimization for second-order multi-agent systems with dynamic event-triggered communication. In *IEEE Conference on Decision and Control*, 3397–3402.

Yi, X., Zhang, S., Yang, T., Chai, T., and Johansson, K.H. (2019). Exponential convergence for distributed smooth optimization under the restricted secant inequality condition. *arXiv preprint arXiv:1909.03282*.

Yu, W., Yi, P., and Hong, Y. (2016). A gradient-based dissipative continuous-time algorithm for distributed optimization. In *Chinese Control Conference*, 7908–7912.

Yuan, K., Ling, Q., and Yin, W. (2015). On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3), 1835–1854.

Yuan, Y., Tang, X., Zhou, W., Pan, W., Li, X., Zhang, H., Ding, H., and Goncalves, J. (2019). Data driven discovery of cyber physical systems. *Nature communications*, 10(4894).

Zeng, J. and Yin, W. (2017). Extrapush for convex smooth decentralized optimization over directed networks. *Journal of Computational Mathematics*, 35(4), 383–396.

Zhang, H. and Cheng, L. (2015). Restricted strong convexity and its applications to convergence analysis of gradient-type methods in convex optimization. *Optimization Letters*, 9(5), 961–979.

Zhang, Y., Deng, Z., and Hong, Y. (2017). Distributed optimal coordination for multiple heterogeneous Euler–Lagrangian systems. *Automatica*, 79, 207–213.

Zhu, M. and Martínez, S. (2011). On distributed convex optimization under inequality and equality constraints. *IEEE Transactions on Automatic Control*, 57(1), 151–164.