Abstract—We develop a fast distributed algorithm, termed DEEXTRA, to solve optimization problems when $N$ agents reach agreement and collaboratively minimize the sum of their local objectives over the network, where the communications between agents are described by a directed graph. Existing algorithms, including Gradient-Push (GP) and Directed-Distributed Gradient Descent (D-DGD), solve this problem restricted to directed graphs with a convergence rate of $O(\ln k/\sqrt{k})$. Our analysis shows that DEEXTRA converges at a linear rate $O(\epsilon^k)$ for some constant $\epsilon < 1$, with the assumption that the objective functions are strongly convex. Simulation examples illustrate our findings.

I. INTRODUCTION

We consider the distributed multi-agent optimization over directed graphs of minimizing a sum of objectives, $\sum_{i=1}^{n} f_i(x)$, where $f_i : \mathbb{R}^p \rightarrow \mathbb{R}$ is a private objective function at the $i$th agent of the network. When the network topology is described by undirected graphs, there have been significant progress on algorithms solving the problem. To summarize, the related algorithms can be related to Distributed Gradient Descent (DGD), [1], or the distributed implementations of the Alternating Direction Method of Multipliers (ADMM), [2]. The gradient-based methods have the advantage of computational simplicity, but suffer from a slow convergence rate. The distributed implementations of ADMM is fast, at the cost of the high computation burden. Many methods are proposed to overcome the slow convergence rate of DGD, including DGD with constant stepsize, [3], the Distributed Nesterov Gradient (DNG), [4], etc. To resolve the issue of high computation burden of ADMM, improved algorithms are proposed, including the Decentralized Linear ADMM (DLM), [5], and EXTRA, [6]. We skip the detailed description of above algorithms given space limitation. Detailed exploration and comparison of these algorithms can be found in [7].

We report the papers considering directed graphs here. There are two types of alternatives, which are both gradient-based algorithms, solving the problems restricted to directed graphs. One type of algorithms, called Gradient-Push (GP) method, [8] [9], combines gradient descent and push-sum consensus, [10] [11]. The idea is based on computing the stationary distribution of the Markov chain characterized by the multi-agent network and canceling the imbalance between agents caused by directed graphs by dividing with the left-eigenvector. Directed-Distributed Gradient Descent (D-DGD), [12] [13], is the other gradient-based alternative to solve the problem. The main idea follows Cai’s paper, [14], where a non-doubly-stochastic matrix is constructed to reach average-consensus. In each iteration, D-DGD simultaneously constructs a row-stochastic matrix and a column-stochastic matrix instead of only a doubly-stochastic matrix. Since both GP and D-DGD are gradient-based algorithms, they suffer from a slow convergence rate of $O(\ln k/\sqrt{k})$ due to the diminishing stepsize.

In this paper, we propose a fast distributed algorithm, termed DEEXTRA, to solve the problem over directed graphs, by combining the push-sum protocol and EXTRA, [6]. The push-sum protocol has proven to be a good technique for dealing with optimization over digraphs, [8] [9]. EXTRA works well in optimization problems over undirected graph with a fast convergence rate and a low computation complexity. By applying the push-sum techniques into EXTRA, we show that DEEXTRA converges exactly to the optimal solution with a linear rate, $O(\epsilon^k)$, even when the underlying network is directed. We note that currently our formulation is restricted to strongly convex functions.

The remainder of the paper is organized as follows. Section II describes, and interprets the DEEXTRA algorithm. Section III presents the appropriate assumptions and states the main convergence results. We show numerical results in Section IV, and Section V contains the concluding remarks.

Notation: We use lowercase bold letters to denote vectors and uppercase italic letters to denote matrices. We denote by $[x]_i$ the $i$th component of $x$. For matrix $A$, we denote by $[A]_{ij}$ the $i$th row of the matrix, and by $[A]_{ij}$ its $(i,j)$th element. The matrix, $I_n$, represents the $n \times n$ identity, and $1_n$ ($0_n$) are the $n$-dimensional vector of all 1’s (0’s). The inner product of two vectors $x$ and $y$ is $(x,y)$. We define the $A$-matrix norm, $\|x\|_{A}^2 = \langle x,Ax \rangle$ for $A+A^\top$ being positive semi-definite. If a symmetric matrix $A$ is positive semi-definite, we write $A \succeq 0$, while $A \succ 0$ means $A$ is positive definite. The largest and smallest eigenvalues of a matrix $A$ are denoted as $\lambda_{\text{max}}(A)$ and $\lambda_{\text{min}}(A)$. The smallest non-zero eigenvalue of a matrix $A$ is denoted as $\lambda_{\text{min}}(A)$.
Define $\mathcal{N}^{\text{in}}_i$ to be the collection of in-neighbors, i.e., the set of agents that can send information to agent $i$. Similarly, $\mathcal{N}^{\text{out}}_i$ is defined as the out-neighborhood of agent $i$, i.e., the set of agents that can receive information from agent $i$. We allow both $\mathcal{N}^{\text{in}}_i$ and $\mathcal{N}^{\text{out}}_i$ to include the node $i$ itself. Note that $\mathcal{N}^{\text{in}}_i \neq \mathcal{N}^{\text{out}}_i$, in general. We focus on solving an optimization problem that is distributed over the above multi-agent network as follows:

$$P_1: \quad \min f(x) = \sum_{i=1}^{n} f_i(x),$$

where each $f_i : \mathbb{R}^p \rightarrow \mathbb{R}$ is only known by agent $i$.

### A. EXTRA for undirected graphs

EXTRA is a fast exact first-order algorithm to solve Problem P1 but in the case that the communication network is described by an undirected graph. At the $k$th iteration, each agent $i$ performs the following update:

$$x^{k+1}_i = x^k_i + \sum_{j \in \mathcal{N}_i} w_{ij} x^k_j - \sum_{j \in \mathcal{N}_i} \bar{w}_{ij} x^{k-1}_j - \alpha \left[ \nabla f_i(x^k_i) - \nabla f_i(x^{k-1}_i) \right], \quad (1)$$

where the weight, $w_{ij}$, forms a weighting matrix, $W = \{w_{ij}\}$ satisfying that $W$ is symmetric and doubly-stochastic, and the collection $\bar{W} = \{\bar{w}_{ij}\}$ satisfies $\bar{W} = \theta \mathbf{I}_n + (1 - \theta)W$ with some $\theta \in (0, \frac{1}{2}]$. The update in Eq. (1) ensures $x^k_i$ to converge to the optimal solution of the problem for all $i$ with convergence rate $O(\frac{1}{k})$, and converge linearly when the objective function is strongly-convex. To better represent EXTRA and later compare with DEXTRA, we write Eq. (1) in a matrix form. Let $x^k, \nabla f(x^k) \in \mathbb{R}^{np}$ be the collections of all agents states and gradients at time $k$, i.e., $x^k \triangleq [x^k_1; \ldots; x^k_n], \nabla f(x^k) \triangleq [\nabla f_1(x^k_1); \ldots; \nabla f_n(x^k_n)],$ and $W, \bar{W} \in \mathbb{R}^{nxn}$ be the weighting matrices collecting weights, $w_{ij}, \bar{w}_{ij}$, respectively. Then Eq. (1) can be represented in a matrix form as:

$$x^{k+1} = (I_n + W \otimes I_p)x^k - (\bar{W} \otimes I_p)x^{k-1} - \alpha \left[ \nabla f(x^k) - \nabla f(x^{k-1}) \right], \quad (2)$$

where the symbol $\otimes$ is the Kronecker product. We now state DEXTRA and derive it in a similar form as EXTRA.

### B. DEXTRA Algorithm

To solve the Problem P1 suited to the case of directed graphs, we claim the DEXTRA algorithm implemented in a distributed manner as follow. Each agent, $j \in \mathcal{V}$, maintains two vector variables: $x^k_j, \dot{x}^k_j \in \mathbb{R}^p$, as well as a scalar variable, $y^k_j \in \mathbb{R}$, where $k$ is the discrete-time index. At the $k$th iteration, agent $j$ weights its state estimate, $a_{ij} x^k_j$, $a_{ij} \dot{x}^k_j$, and $\bar{a}_{ij} x^{k-1}_j$, and sends it to each out-neighbor, $i \in \mathcal{N}^{\text{out}}_j$, where the weights, $a_{ij}$ and, $\bar{a}_{ij}$, $s$ are such that:

$$a_{ij} = \begin{cases} 
0, & i \in \mathcal{N}^{\text{out}}_j, \\
\frac{1}{n}, & \text{otherwise.} 
\end{cases}$$

where $\theta \in (0, \frac{1}{2}]$. With agent $i$ receiving the information from its in-neighbors, $j \in \mathcal{N}^{\text{in}}_i$, it calculates the states, $x^k_i$, by dividing $x^k_i$ over $y^k_i$, and updates the variables $x^k_i$, and $y^{k+1}_i$ as follows:

$$x^k_i = \frac{x^k_i}{y^k_i}, \quad (3a)$$

$$x^{k+1}_i = x^k_i + \sum_{j \in \mathcal{N}^{\text{in}}_i} (a_{ij} x^k_j) - \sum_{j \in \mathcal{N}^{\text{out}}_i} (\bar{a}_{ij} x^{k-1}_j) - \alpha \left[ \nabla f_i(x^k_i) - \nabla f_i(x^{k-1}_i) \right], \quad (3b)$$

$$y^{k+1}_i = \sum_{j \in \mathcal{N}^{\text{in}}_i} (a_{ij} y^k_j), \quad (3c)$$

where $\nabla f_i(x^k_i)$ is the gradient of the function $f_i(z)$ at $z = z^k_i$, and $\nabla f_i(x^{k-1}_i)$ is the gradient at $z^{k-1}_i$, respectively. The method is initiated with an arbitrary vector, $x^0_i$, and with $y^0_i = 1$ for any agent $i$. The stepsize, $\alpha$, is a positive number within a certain interval. We will explicitly discuss the range of $\alpha$ in Section III as well as in the numerical experiments in Section IV. At the first iteration, i.e., $k = 0$, we have instead the iteration $x^1_i = \sum_{j \in \mathcal{N}^{\text{in}}_i} (a_{ij} x^0_j) - \alpha \nabla f_i(x^0_i)$.

To simplify the proof, we write DEXTRA, Eq. (3), in a matrix form. Let, $A = \{a_{ij}\}, \bar{A} = \{\bar{a}_{ij}\}$, be the collection of weights, $a_{ij}, \bar{a}_{ij}$, respectively, $x^k, \dot{x}^k, \nabla f(x^k) \in \mathbb{R}^{np}$, be the collection of all agents states and gradients at time $k$, i.e., $x^k \triangleq [x^k_1; \ldots; x^k_n], \dot{x}^k \triangleq [\dot{x}^k_1; \ldots; \dot{x}^k_n], \nabla f(x^k) \triangleq [\nabla f_1(x^k_1); \ldots; \nabla f_n(x^k_n)],$ and $W, \bar{W} \in \mathbb{R}^{nxn}$ be the collection of agent states, $y^k, \dot{y}^k$, i.e., $y^k \triangleq [y^k_1; \ldots; y^k_n]$. Note that at any time $k$, $y^k$ can be represented by the initial value, $y^0$:

$$y^k = A y^{k-1} + A^k y^0 = A^k \cdot 1_n. \quad (4)$$

Define the diagonal matrix $D^k$ for any $k$ such that the $i$th element of $D^k$ is $y^k_i$, i.e.,

$$D^k = \text{diag}(A^k \cdot 1_n). \quad (5)$$

Then we have Eq. (3) in the matrix form equivalently as follows:

$$x^{k+1} = (D^k \otimes I_p) x^k, \quad (6a)$$

$$x^{k+1} = x^k + (A \otimes I_p) x^k - (\bar{A} \otimes I_p) x^{k-1} - \alpha \left[ \nabla f(x^k) - \nabla f(x^{k-1}) \right], \quad (6b)$$

$$y^{k+1} = A y^k, \quad (6c)$$

where both the two weight matrices, $A$ and $\bar{A}$, are column stochastic matrix and satisfy the relationship: $\bar{A} = \theta I_n + (1 - \theta) A$ with some $\theta \in (0, \frac{1}{2}]$. When $k = 0$, Eq. (6b) changes to $x^1 = (A \otimes I_p)x^0 - \alpha \nabla f(x^0)$. Consider Eq. (6a), we obtain

$$x^k = (D^k \otimes I_p) z^k, \quad \forall k. \quad (7)$$

Therefore, by substituting $x^k$ into Eq. (6b), we are possible to represent DEXTRA in a similar form as EXTRA:

$$x^{k+1} = (D^k \otimes I_p) z^{k+1} = [(I_n + A) D^k \otimes I_p] z^k$$
\[- (\bar{A} D^{k-1} \otimes I_p) z^{k-1} - \alpha \left[ \nabla f(z^k) - \nabla f(z^{k-1}) \right]. \]  (8)

C. Interpretation of DEXTRA

In this section, we give an intuitive interpretation on DEXTRA’s convergence to the optimal solution before we show the main result. Note that the analysis given here is not a formal proof of DEXTRA. However, it gives a quick snapshot on how DEXTRA works.

Since $A$ is a column stochastic matrix, the sequence of iterates, $\{y^k\}$, generated by Eq. (5c) converges to the span of $A$’s right eigenvector corresponding to eigenvalue 1, i.e., $y^\infty = u \cdot \pi$ for some number $u$, where $\pi$ is the right eigenvector of $A$ corresponding to eigenvalue 1. We assume that the sequence of iterates, $\{z^k\}$ and $\{x^k\}$, generated by DEXTRA, Eq. (6) or (8), converges to their own limit points, $z^\infty$, $x^\infty$, respectively, (which might not be the truth).

According to the updating rule in Eq. (6b), the limit point $z^\infty$ satisfies

$$z^\infty = (D^\infty)^{-1} \otimes I_p \left( \pi \otimes v \right) \in \{1_n \otimes 0\},$$  (10)

where the consensus is reached. The above analysis reveals the idea of DEXTRA to overcome the imbalance of agent states occurred when the graph is directed: both $x^\infty$ and $y^\infty$ lie in the span of $\pi$. By dividing $x^\infty$ over $y^\infty$, the imbalance is canceled.

By summing up the updates in Eq. (6b) over $k$, we obtain that

$$x^{k+1} = (A \otimes I_p) x^k - \alpha \nabla f(z^k) - \sum_{r=0}^{k-1} \left[ (\bar{A} - A) \otimes I_p \right] x^r.$$  (11)

Consider that $x^\infty = \pi \otimes v$ and the preceding relation. It follows that the limit point $z^\infty$ satisfies

$$\alpha \nabla f(z^\infty) = - \sum_{r=0}^{\infty} \left[ (\bar{A} - A) \otimes I_p \right] x^r.$$  (12)

Therefore, we obtain that

$$\alpha (1_n \otimes I_p) \nabla f(z^\infty) = - \left[ 1_n^\top (\bar{A} - A) \otimes I_p \right] \sum_{r=0}^{\infty} x^r = 0_p,$$

which is the optimality condition of Problem P1. Therefore, given the assumption that the sequence of DEXTRA iterates, $\{z^k\}$ and $\{x^k\}$, have limit points, $z^\infty$ and $x^\infty$, the limit point, $z^\infty$, achieves consensus and reaches the optimal solution of Problem P1. In next section, we state our main result of this paper.

III. ASSUMPTIONS AND MAIN RESULTS

With appropriate assumptions, our main result states that DEXTRA converges to the optimal solution of Problem P1 linearly. In this paper, we assume that the agent graph, $G$, is strongly-connected; each local function, $f_i : \mathbb{R}^p \rightarrow \mathbb{R}$, is convex and differentiable, $\forall i \in V$, and the optimal solution of Problem P1 and the corresponding optimal value exist. Formally, we denote the optimal solution by $\phi \in \mathbb{R}^p$ and optimal value by $f^*$, i.e.,

$$f(\phi) = f^* = \min_{x \in \mathbb{R}^p} f(x).$$  (13)

Besides the above assumptions, we emphasize some other assumptions regarding the objective functions and weighting matrices, which are formally presented as follow.

Assumption A1 (Functions and Gradients). For all agent $i$, each function, $f_i$, is differentiable and satisfies the following assumptions.

(a) The function, $f_i$, has Lipschitz gradient with the constant $L_{f_i}$, i.e., $\| \nabla f_i(x) - \nabla f_i(y) \| \leq L_{f_i} \| x - y \|$, $\forall x, y \in \mathbb{R}^p$.

(b) The gradient, $\nabla f_i(x)$, is bounded, i.e., $\| \nabla f_i(x) \| \leq B_{f_i}, \forall x \in \mathbb{R}^p$.

(c) The function, $f_i$, is strictly convex with a constant $S_{f_i}$, i.e., $S_{f_i} \| x - y \|^2 \leq \langle \nabla f_i(x) - \nabla f_i(y), x - y \rangle$, $\forall x, y \in \mathbb{R}^p$.

Following Assumption A1, we have for any $x, y \in \mathbb{R}^p$,

$$\| \nabla f(x) - \nabla f(y) \| \leq L_{f_i} \| x - y \|,$$  (14a)

$$\| \nabla f(x) \| \leq B_{f_i},$$  (14b)

$$S_{f_i} \| x - y \|^2 \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle,$$  (14c)

where the constants $B_{f_i} = \max_i \{B_{f_i}\}$, $L_{f_i} = \max_i \{L_{f_i}\}$, and $S_{f_i} = \max_i \{S_{f_i}\}$. By combining Eqs. (14a) and (14c), we obtain

$$\| x - y \| \leq \frac{1}{S_{f_i}} \| \nabla f(x) - \nabla f(y) \| \leq \frac{2B_{f_i}}{S_{f_i}},$$

which is

$$\| x \| \leq \frac{2B_{f_i}}{S_{f_i}},$$  (15)

by letting $y = 0_{np}$. Eq. (14) reveals that the sequences updated by DEXTRA is bounded.

Recall the definition of $D^k$ in Eq. (5), we denote the limits of $D^k$ by $D^\infty$, i.e.,

$$D^\infty = \lim_{k \rightarrow \infty} D^k = \text{diag} (A^\infty \cdot 1_n) = n \cdot \text{diag} (\pi),$$  (16)

where $\pi$ is the right eigenvector of $A$ corresponding to eigenvalue 1. The next assumption is related to the weighting matrices, $A$ and $\bar{A}$, and $D^\infty$.

Assumption A2 (Weighting matrices). The weighting matrices, $A$ and $\bar{A}$, used in DEXTRA, Eq. (6) or (8), satisfy the following.

(a) $A$ is a column stochastic matrix.
(b) $\tilde{A}$ is a column stochastic matrix and satisfies $\tilde{A} = \theta I_n + (1 - \theta) A$, for some $\theta \in (0, \frac{1}{2}]$.

(c) $(D^\infty)^{-1} \tilde{A} + \tilde{A}^\top (D^\infty)^{-1} > 0$.

To ensure $(D^\infty)^{-1} \tilde{A} + \tilde{A}^\top (D^\infty)^{-1} > 0$, the agent $i$ can assign a large value to its own weight $a_{ii}$ in the implementation of DEXTRA, (such that $A$ is more like a diagonally dominant matrix). Then all eigenvalues of $(D^\infty)^{-1} \tilde{A} + \tilde{A}^\top (D^\infty)^{-1}$ can be positive. It is also important in this paper to show that $(D^\infty)^{-1} (\tilde{A} - A) + (\tilde{A} - A)^\top (D^\infty)^{-1} \geq 0$ to help establish the convergence of DEXTRA. We introduce the following lemma.

Lemma 1. (Chung [15]) The Laplacian, $L$, of a directed graph, $G$, is defined by

$$L(G) = I_n - \frac{S^{1/2}RS^{-1/2} + S^{-1/2}R^\top S^{1/2}}{2},$$

where $R$ is the transition probability matrix and $S = \text{diag}(s)$, where $s$ is the left eigenvector of $R$ corresponding to eigenvalue 1. If $G$ is strongly connected, then the eigenvalues of $L(G)$ satisfy $0 = \lambda_0 < \lambda_1 < \cdots < \lambda_n$.

Since

$$(D^\infty)^{-1} (\tilde{A} - A) + (\tilde{A} - A)^\top (D^\infty)^{-1} = \theta (D^\infty)^{-1/2} L(G) (D^\infty)^{-1/2},$$

it follows from Lemma 1 that

$$\lambda_{\min} \left( (D^\infty)^{-1} (\tilde{A} - A) + (\tilde{A} - A)^\top (D^\infty)^{-1} \right) = 0,$$

and for any $a \in \mathbb{R}^n$, we have

$$\|a\|^2_{(D^\infty)^{-1}(\tilde{A} - A)} \geq 0, \quad \forall a \in \mathbb{R}^n.$$  

Before finally stating the main result, we define some auxiliary sequences. Let $z^* \in \mathbb{R}^{np}$ be the collection of optimal solution, $\phi$, of Problem P1, i.e.,

$$z^* = 1_n \otimes \phi,$$

and $q^* \in \mathbb{R}^{np}$ some vector satisfying

$$\left(\tilde{A} - A\right)^\top I_p \ q^* + \alpha \nabla f(z^*) = 0_{np}. $$

Introduce auxiliary sequence

$$\tilde{q}^k = \sum_{r=0}^{k} x^r,$$

and for each $k$,

$$t^k = \left((D^\infty \otimes I_p) z^k\right), \quad t^* = \left((D^\infty \otimes I_p) z^*\right),$$

$$G = \left[\begin{array}{c}
(\tilde{A}^\top (D^\infty)^{-1}) \otimes I_p \\
((D^\infty)^{-1} (\tilde{A} - A)) \otimes I_p \end{array}\right].$$

In order to state the theorem with simplified representations, we also need the following notation. Denote $P = I_n - A$, $L = \tilde{A} - A$, $N = (D^\infty)^{-1} (\tilde{A} - A)$, $M = (D^\infty)^{-1} \tilde{A}$, $d_{\max} = \max_k \{\|D^k\|\}$, $d_{\max}^- = \max_k \{\|(D^k)^{-1}\|\}$, and $d_{\infty}^- = \|(D^\infty)^{-1}\|$. Let $C_1 = \left(\frac{\lambda_{\min}(NN^\top) + \lambda_{\max}(NN^\top)}{2}\right)^\frac{1}{2}$, $C_2 = 8C_1 (L_d d_{\max})^2$, $C_3 = \lambda_{\max} \left(\frac{M + M^\top}{2}\right) + 4C_1 (1 - 2\theta)^2 P^\top P$, and $\Delta = \left(\frac{S_f}{2d_{\max}} - \frac{\eta}{2}\right)^2 - 4C_1 \delta \left(\frac{\eta}{2} + C_3 \delta\right)$. We are now ready to state the main result as shown in Theorem 1 which makes explicit the rate at which the objective function converges to its optimal value.

Theorem 1. Let Assumptions [A1] [A2] hold. Then with proper step size $\alpha \in [\alpha_{\min}, \alpha_{\max}]$, there exists, $\delta > 0$, $|\Gamma| < \infty$, and $\gamma(0, 1)$, such that the sequence $\{t^k\}$ generated by DEXTRA satisfies

$$\|t^k - t^*\|^2_G \geq (1 + \delta) \|t^{k+1} - t^*\|^2_G - \Gamma \gamma^k.$$  

The lower bound, $\alpha_{\min}$, of $\alpha$ satisfies

$$\alpha_{\min} \leq \alpha \leq \frac{S_f}{2d_{\max}} - \frac{\eta}{2} - \sqrt{\Delta},$$

and the upper bound, $\alpha_{\max}$ of $\alpha$ satisfies

$$\alpha_{\max} \geq \alpha \geq \frac{\eta \lambda_{\min}(M + M^\top)}{2} \frac{S_f}{2d_{\max}} - \frac{\eta}{2} + \sqrt{\Delta},$$

where, $\eta$, is some arbitrary constant to ensure the solution set of $\alpha > 0$ is nonempty and contains positive values.

Proof. The proof can be found in [7]. \hfill \Box

Note that the result of Theorem 1 Eq. (22), does not equivalently mean that $t^k \rightarrow t^*$. To prove $t^k \rightarrow t^*$, we further need to show that the matrix $G$ satisfies $\|t^k - t^*\|^2_G \geq 0, \forall k$. It follows from the definition of $t^k, t^*$ and $G$, Eq. (21), that

$$\|t^k - t^*\|^2_G = \left( (D^\infty \otimes I_p) z^k - (D^\infty \otimes I_p) z^* \right)^\top (\tilde{A}^\top (D^\infty)^{-1}) \otimes I_p (D^\infty \otimes I_p) (\tilde{A}^\top (D^\infty)^{-1}) \otimes I_p^\top (D^\infty \otimes I_p) (\tilde{A}^\top (D^\infty)^{-1}) \otimes I_p^\top (D^\infty \otimes I_p) (\tilde{A}^\top (D^\infty)^{-1}) \otimes I_p.$$ 

With Assumption [A2] (c), we already have $\|a\|^2_{(D^\infty)^{-1} \tilde{A}^\top (D^\infty)^{-1}} \geq 0, \forall a \in \mathbb{R}^n$. The fact that $\|a\|^2_{(D^\infty)^{-1} \tilde{A}^\top (D^\infty)^{-1}} \geq 0, \forall a \in \mathbb{R}^n$ is shown in Eq. (17).

Therefore, $\|t^k - t^*\|^2_G$ converges to 0 at the R-linear rate $O((1 + \delta)^{-k})$ for some $\delta \in (0, \delta)$. Consequently, both $\|D^k z^k - D^\infty z^*\|^2_{\tilde{A}^\top (D^\infty)^{-1}} \otimes I_p$ and $\|q^k - q^*\|^2_{(D^\infty)^{-1} (\tilde{A} - A) \otimes I_p}$ converge to 0 at the R-linear rate $O((1 + \delta)^{-k})$ for some $\delta \in (0, \delta)$. Again with Assumption [A2] (c), it follows that $\tilde{z}^k \rightarrow z^*$ with a R-linear rate, i.e., the sequence $\{z^k\}$, generated by DEXTRA converges to the optimal solution of the problem.

\footnote{We note that $\|a\|^2_{\tilde{A}^\top (D^\infty)^{-1}} \geq 0$ as long as $\{q^k - q^*\} \in \pi$, where $\pi$ is the right eigenvector of $A$ corresponding to eigenvalue 1. In fact, with the definition of $q^k$, Eq. (20), and $q^*$, Eq. (19), we have $\|q^k - q^*\|^2 \neq 0$, but $(q^k - q^*) \in \text{span}(\pi)$.}
Theorem 1 gives an specific value of bounds on $\alpha_{\text{min}}$ and $\alpha_{\text{max}}$. In order to ensure that the solution set of stepsize, $\alpha$, is not empty, i.e., $\alpha_{\text{min}} \leq \alpha_{\text{max}}$, it is sufficient to satisfy
\[
\Delta = \left( \frac{S_f}{2d_{\text{max}}} - \frac{\eta}{2} \right)^2 - 4C_3\delta \left( \frac{1}{\delta} + C_4\delta \right) \geq 0, \quad \tag{24}
\]
\[
\alpha = \frac{S_f}{2d_{\text{max}}} - \frac{\eta}{2} - \sqrt{\Delta} \leq \frac{\eta\lambda_{\text{min}}(M + M^\top)}{2L_f(d_{\text{max}}^2 - d_{\text{max}}^2)} = \bar{\alpha}. \quad \tag{25}
\]

Eqs. (24) and (25) set some restrictions on the strongly-convexity constant, $S_f$. In real situation, the linear convergence rate is obtained for any strongly convex functions. The theoretical restriction here is due to the bounds of stepsize we give in Theorem 1 not being perfectly tight, i.e., $\alpha_{\text{min}} < \alpha$, $\alpha_{\text{max}} > \bar{\alpha}$. However, the explicit value of $\alpha$ and $\bar{\alpha}$ implies how to increase the interval of stepsize, therefore decreasing the restriction on $S_f$. For example, it is wise to set the weights to increase $\lambda_{\text{min}}(M + N^\top)/d_{\text{max}}^2$ such that $\bar{\alpha}$ is increased. We will show those properties in numerical experiments in Section IV.

IV. Numerical Experiments

This section provides numerical experiments to study the convergence rate of DEXTRA for a least squares problem in a directed graph. The distributed least squares problem is described as follows. Each agent owns a private objective function, $s_i = R_i x + n_i$, where $s_i \in \mathbb{R}^{m_i}$ and $R_i \in \mathbb{R}^{m_i \times p}$ are measured data, $x \in \mathbb{R}^p$ is the unknown, and $n_i \in \mathbb{R}^{m_i}$ is random noise. The goal is to estimate $x$, which we formulate as a distributed optimization problem solving
\[
\min f(x) = \frac{1}{n} \sum_{i=1}^n \| R_i x - s_i \|. \quad \tag{26}
\]

We consider the network topology as the digraph shown in Fig. 1. The default weighting strategy used is the local degree weighting strategy: to assign each agent itself and its out-neighbors equal weights according to agent’s own out-degree, i.e.,
\[
a_{ij} = \frac{1}{\delta_i}, \quad (i, j) \in \mathcal{E}. \quad \tag{26}
\]

According to this weighting strategy, the corresponding network parameters are shown in Fig. 1. Recall the representation of $\alpha$ and $\bar{\alpha}$ given in Theorem 1, we have $\bar{\alpha} = 0.26$, and $\alpha = 5 \times 10^{-4}$. We thus pick $\alpha = 0.1 \in [\alpha, \bar{\alpha}]$ for the following experiments.

Our first experiment compares several algorithms suited to directed graphs, illustrated in Fig. 1. The comparison of DEXTRA, GP, D-DGD and DGD with weighting matrix being row-stochastic is shown in Fig. 2. The convergence rate of DEXTRA is linear as stated in Section III. G-P and D-DGD apply the same stepsize, $\alpha = \frac{\eta}{\bar{\alpha}}$. As a result, the convergence rate of both is sub-linear. We also consider the DGD algorithm, but with the weighting matrix being row-stochastic. The reason is that in a directed graph, it is impossible to construct a doubly-stochastic matrix. As expected, DGD with row-stochastic matrix does not converge to the exact optimal solution while other three algorithms are suited to the case of directed graph.

According to the theoretical value of $\alpha$ and $\bar{\alpha}$, we are able to set available stepsize, $\alpha \in [5 \times 10^{-4}, 0.26]$. In practice, this interval is much wider. Fig. 3 illustrates this fact. Numerical experiments show that $\alpha_{\text{min}} = 0$ and $\alpha_{\text{max}} = 0.447$.

The explicit representation of $\bar{\alpha}$ and $\alpha$ given in Theorem 1 imply the way to increase the interval of stepsize, i.e.,
\[
\bar{\alpha} \propto \frac{\lambda_{\text{min}}(M + M^\top)}{(d_{\text{max}}d_{\text{max}}^2)^2}, \quad \alpha \propto \frac{1}{(d_{\text{max}}d_{\text{max}}^2)^2}.
\]

To increase $\bar{\alpha}$, we increase $\lambda_{\text{min}}(M + M^\top)/(d_{\text{max}}d_{\text{max}}^2)^2$; to decrease $\alpha$, we can decrease $1/(d_{\text{max}}d_{\text{max}}^2)^2$. Compared with applying the local degree weighting strategy, Eq. (26), as shown in Fig. 3, it can be seen that a a wider range of stepsize can we use while we apply the constant weighting strategy, which can be expressed as
\[
a_{ij} = \begin{cases} 1 - 0.01\delta_i, & i = j, \\ 0.01, & (i, j) \in \mathcal{E}. \end{cases} \quad \tag{27}
\]
This constant weighting strategy constructs a “diagonal-dominant” weighting matrix, which increases \( \lambda_{\text{min}}(M + M^\top) \). It may be observed from Figs. 3 and 4 that the same stepsize generates quiet different convergence speed when the weighting strategy changes. Compared Figs. 3 and 4 when stepsize \( \alpha = 0.1 \), DEXTRA with local degree weighting converges much faster.

V. CONCLUSIONS

We introduce DEXTRA, a distributed algorithm to solve multi-agent optimization problems over directed graphs. We have shown that DEXTRA succeeds in driving all agents to the same point, which is the optimal solution of the problem, given that the communication graph is strongly-connected and the objective functions are strongly convex. Moreover, the algorithm converges at a linear rate \( O(\epsilon^k) \) for some constant \( \epsilon < 1 \). The constant depends on the network topology \( G \), as well as the Lipschitz constant, strong convexity constant, and gradient norm of objective functions. Numerical experiments are conducted for a least squares problem, where we show that DEXTRA is the fastest distributed algorithm among all algorithms suited to the case of directed graphs, compared with GP, and D-DGD.

REFERENCES