IEEE
CONTROL SYSTEMS LETTERS

IEEE Xplore® Digital Library
http://ieeexplore.ieee.org/Xplore

A linear algorithm for optimization over directed graphs with geometric convergence

Ran Xin, *Student Member, IEEE*, and Usman A. Khan, *Senior Member, IEEE*

***Abstract*—In this letter, we study distributed optimization, where a network of agents, abstracted as a directed graph, collaborates to minimize the average of locally-known convex functions. Most of the existing approaches over directed graphs are based on push-sum (type) techniques, which use an independent algorithm to asymptotically learn either the left or right eigenvector of the underlying weight matrices. This strategy causes additional computation, communication, and nonlinearity in the algorithm. In contrast, we propose a *linear* algorithm based on an inexact gradient method and a gradient estimation technique. Under the assumptions that each local function is strongly-convex with Lipschitz-continuous gradients, we show that the proposed algorithm geometrically converges to the global minimizer with a sufficiently small step-size. We present simulations to illustrate the theoretical findings.**

***Index Terms*—Distributed optimization, directed graphs**

I. Introduction

In this letter, we consider distributed optimization over multi-agent networks. Formally, each agent i has access only to a private function, $f_i : \mathbb{R}^p \to \mathbb{R}$. The goal is to minimize the average of these functions, $\frac{1}{n}\sum_{i=1}^n f_i(\mathbf{x})$, via information exchange among the agents. We focus on the case where the communication network is described by an arbitrary *directed* graph. Early work on distributed optimization includes distributed sub-gradient descent (DGD) [1], which converges to the optimal solution at a sublinear rate, i.e., $O(\frac{\ln k}{\sqrt{k}})$ for arbitrary (possibly non-differentiable) convex functions and $O(\frac{\ln k}{k})$ for strongly-convex functions, where k is the number of iterations. These methods are slow due to the diminishing step-sizes. With the help of strong-convexity and Liptschiz-continuous gradients, algorithms with faster convergence rates have been developed. In particular, DGD with a constant step-size [2] converges geometrically to an error ball around the optimal solution. Another method, EXTRA [3], achieves geometric convergence to the global optimal solution with the requirement of symmetric weights. Of relevance are Refs. [4]–[6], which combine inexact gradient methods and a gradient estimation technique based on dynamic average consensus [7]. Additional related work and applications can be found in [8]–[13].

All of the aforementioned methods require the underlying graphs to be undirected or weight-balanced. This requirement, however, may not be practical, for example, when the agents broadcast at different power levels leading to communication capability in one direction but not in the other. It is natural thus to develop optimization and learning algorithms that are applicable to directed graphs. The primary challenge in dealing with directed graphs is that it may not be possible to construct doubly-stochastic weight matrices for information fusion. The weighted adjacency matrix for directed graphs, in general, may only be either row-stochastic or column-stochastic, but not both. See [14] for work on balancing the weights in strongly-connected directed graphs.

The existing approaches for optimization over directed graphs are motivated by combining average-consensus methods developed for directed graphs with optimization algorithms designed for undirected graphs. For instance, subgradient-push introduced in [15] and further studied in [16] combines push-sum consensus [17] and DGD; A linear algorithm over directed graphs, called Directed-Distributed Gradient Descent (D-DGD), was introduced in [18], [19], and is based on surplus consensus [20] and DGD. Such DGD-based methods, however, restricted by the diminishing step-size, converge relatively slowly at $O(\frac{\ln k}{\sqrt{k}})$ for general convex functions and $O(\frac{\ln k}{k})$ for strongly-convex functions. The convergence rate has been recently improved in DEXTRA [21], which converges geometrically to the global optimal given that its step-size lies in an interval and the objective functions are strongly-convex with Lipschitz-continuous gradients. DEXTRA was subsequently improved in ADD-OPT/Push-DIGing [22], [23], which geometrically converges with a sufficiently small step-size. The implementation of DEXTRA and ADD-OPT/Push-DIGing requires each agent to know its out-degree in order to construct a column-stochastic weight matrix. This requirement is later removed in [24] and FROST [25], which use row-stochastic weights and thus require no knowledge of out-degrees as each agent locally decides weights assigned to the incoming information. What is common among these fast methods over directed graphs is that they all are based on push-sum (type) techniques, which make the resulting algorithm nonlinear because an independent algorithm is used to asymptotically learn either the right or the left eigenvector, corresponding to the eigenvalue of 1, of the weight matrix. This strategy causes additional computation and communication on the agents.

In this paper, we provide a *linear* distributed optimization algorithm that converges geometrically to the global optimal with a sufficiently small step-size and when the objective functions are strongly-convex with Lipschitz-continuous gradients. The analysis rests on a novel matrix norm argument and a matrix perturbation result. In the rest of the paper, Section II provides the algorithm development and its relationship with existing approaches, while Section III details the convergence

R. Xin and U. A. Khan are with the Department of Electrical and Computer Engineering, Tufts University, 161 College Ave, Medford, MA 02155; ran.xin@tufts.edu, khan@ece.tufts.edu. This work has been partially supported by an NSF Career Award # CCF-1350264.

Digital Object Identifier: 10.1109/LCSYS.2018.2834316

analysis. Section IV presents numerical experiments and Section V concludes the paper.

Basic Notation: We use lowercase bold letters to denote vectors and uppercase italic letters to denote matrices. The matrix, I_n, represents the $n \times n$ identity, whereas $\mathbf{1}_n$ is the n-dimensional column vector of all 1's. For an arbitrary vector, $\mathbf{x}$, we denote its ith element by $[\mathbf{x}]_i$. We denote by $X \otimes Y$, the Kronecker product of two matrices, X and Y. For a matrix, X, we denote $\rho(X)$ as its spectral radius and X_∞ as its infinite power (if it exists), i.e., $X_\infty = \lim_{k\to\infty} X^k$. For a primitive, row-stochastic matrix, $\underline{A}$, we denote its left and right eigenvectors corresponding to the eigenvalue of 1 by $\boldsymbol{\pi}_r$ and $\mathbf{1}_n$, respectively, such that $\boldsymbol{\pi}_r^\top \mathbf{1}_n = 1$. Similarly, for a primitive, column-stochastic matrix, $\underline{B}$, we denote its left and right eigenvectors corresponding to the eigenvalue of 1 by $\mathbf{1}_n$ and $\boldsymbol{\pi}_c$, respectively, such that $\mathbf{1}_n^\top \boldsymbol{\pi}_c = 1$. The notation $\|\cdot\|_2$ denotes the Euclidean norm of vectors and $|||\cdot|||_2$ denotes the spectral norm of matrices.

II. Algorithm Development

In this section, we mathematically formulate the optimization problem and describe the proposed algorithm and its relationship with the existing methods. Consider a network of n agents whose communication links are described by a strongly-connected directed graph, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the index set of agents, and $\mathcal{E}$ is the collection of ordered pairs, $(i, j), i, j \in \mathcal{V}$, such that agent j can send information to agent i, i.e., $j \to i$. We define $\mathcal{N}_i^{\text{in}}$ as the collection of in-neighbors, i.e., the set of agents that can send information to agent i. Similarly, $\mathcal{N}_i^{\text{out}}$ is the set of out-neighbors of agent i. Note that both $\mathcal{N}_i^{\text{in}}$ and $\mathcal{N}_i^{\text{out}}$ include node i. We assume that each agent i knows[1] its out-degree (the number of out-neighbors), denoted by $|\mathcal{N}_i^{\text{out}}|$; see [26] for details.

We focus on solving a convex optimization problem distributed over the above multi-agent network. In particular, the network of agents cooperatively solves the following:

$$\text{P1}: \quad \min\ F(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}),$$

where each $f_i : \mathbb{R}^p \to \mathbb{R}$ is known only to agent i. We assume that each local function, $f_i(\mathbf{x})$, is strongly-convex and has Lipschitz-continuous gradients. Our goal is to design a distributed algorithm such that the iterates at each agent converge to the global optimal solution of Problem P1 via information exchange with nearby agents over the directed graph, $\mathcal{G}$. We formalize the set of assumptions as follows.

Assumption 1. *The graph, $\mathcal{G}$, is strongly-connected and each agent in the network knows its out-degree.*

Assumption 2. *Each local function, f_i, is strongly-convex, and has globally Lipschitz-continuous gradient, i.e., for any i and $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^p$,*

(i) there exists a positive constant β such that
$$\|\nabla f_i(\mathbf{x}_1) - \nabla f_i(\mathbf{x}_2)\|_2 \le \beta \|\mathbf{x}_1 - \mathbf{x}_2\|_2;$$
(ii) there exists a positive constant α such that
$$f_i(\mathbf{x}_1) - f_i(\mathbf{x}_2) \le \nabla f_i(\mathbf{x}_1)^\top(\mathbf{x}_1 - \mathbf{x}_2) - \tfrac{\alpha}{2}\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2.$$
Clearly, the Lipschitz-continuity and strongly-convexity constants for the global objective function $F(\mathbf{x})$ are β and α, respectively. Assumption 2 ensures that the optimal solution, denoted as $\mathbf{x}^*$, for P1 exists and is unique.

Algorithm description: To solve Problem P1, we propose the following algorithm. Each agent, $i \in \mathcal{V}$, maintains two variables: $\mathbf{x}_i(k)$, $\mathbf{y}_i(k) \in \mathbb{R}^p$, where k is discrete-time index. The algorithm, initialized with $\mathbf{y}_i(0) = \nabla f_i(\mathbf{x}_i(0))$ and with arbitrary $\mathbf{x}_i(0), \forall i$, performs the following iterations.

$$\mathbf{x}_i(k+1) = \sum_{j=1}^{n} a_{ij}\mathbf{x}_j(k) - \eta \mathbf{y}_i(k), \tag{1a}$$

$$\mathbf{y}_i(k+1) = \sum_{j=1}^{n} b_{ij}\Big(\mathbf{y}_j(k) + \mathbf{r}_j(k)\Big), \tag{1b}$$

where $\mathbf{r}_j(k) = \nabla f_j\big(\mathbf{x}_j(k+1)\big) - \nabla f_j\big(\mathbf{x}_j(k)\big)$ and the step-size, η, is some positive constant. The weights, a_{ij}'s and b_{ij}'s satisfy the following conditions:

$$a_{ij} = \begin{cases} > 0, & j \in \mathcal{N}_i^{\text{in}}, \\ 0, & \text{otherwise}, \end{cases} \qquad \sum_{j=1}^{n} a_{ij} = 1, \forall i, \tag{2}$$

$$b_{ij} = \begin{cases} > 0, & i \in \mathcal{N}_j^{\text{out}}, \\ 0, & \text{otherwise}, \end{cases} \qquad \sum_{i=1}^{n} b_{ij} = 1, \forall j. \tag{3}$$

Eq. (2) leads to a row-stochastic matrix $\underline{A} = \{a_{ij}\}$, which is easy to implement as each agent locally decides the weights. Eq. (3), on the other hand, results in a column-stochastic matrix $\underline{B} = \{b_{ij}\}$, whose distributed implementation only requires each agent to know its out-degree. In particular, we can construct such weights as $b_{ij} = 1/|\mathcal{N}_j^{\text{out}}|, \forall i, j$.

The algorithm in Eqs. (1) can be explained as follows. To implement Eq. (1a), the receiving agent i decides on the weights a_{ij} assigned to the incoming $\mathbf{x}_j(k)$'s such that a_{ij}'s sum to 1. Implementation of Eq. (1b) requires the sending agent to scale the transmission $\mathbf{y}_j(k) + \mathbf{r}_j(k)$ by appropriate choice of b_{ij}'s (to ensure column-stochasticity of $\underline{B}$) as the out-degree of agent j may not be known to agent i. Agent i subsequently adds these received messages to implement Eq. (1b). Intuitively, Eq. (1b) asymptotically learns the average, $\frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\mathbf{x}_i(k))$, of the local gradients, [4]–[7]; and thus Eq. (1a) approaches a centralized gradient descent, as the descent direction, $\mathbf{y}_i(k)$, becomes the gradient of the global objective function over time.

Relation with existing work: We now briefly compare the proposed algorithm with existing techniques. The algorithms in Refs. [4], [5], can be summarized as a single class of algorithms over undirected graphs with the following form:

$$\mathbf{x}_i(k+1) = \sum_{j=1}^{n} w_{ij}\mathbf{x}_j(k) - \eta \mathbf{y}_i(k), \tag{4a}$$

$$\mathbf{y}_i(k+1) = \sum_{j=1}^{n} w_{ij}\mathbf{y}_j(k) + \mathbf{r}_i(k), \tag{4b}$$

[1]Such an assumption is standard in the related literature, see, e.g., [15], [16], [18]–[22].

where $W = \{w_{ij}\}$ is doubly-stochastic. It is shown in Ref. [5], that Eqs. (4) converge geometrically to the optimal solution of Problem P1 as long as the step-size, η, is sufficiently small. This algorithm, however, is not applicable to directed graphs as it may not be possible to construct doubly-stochastic weights.

To overcome this issue, Refs. [22]–[25] leverage push-sum (type) techniques, with either row- or column-stochastic weights, towards the algorithm in Eqs. (4). Refs. [24], [25], e.g., propose the following algorithm:

$$\begin{aligned}
\mathbf{y}_i(k+1) &= \sum_{j=1}^{n} a_{ij}\mathbf{y}_i(k),\\
\mathbf{x}_i(k+1) &= \sum_{j=1}^{n} a_{ij}\mathbf{x}_i(k) - \eta_i\mathbf{z}_i(k),\\
\mathbf{z}_i(k+1) &= \sum_{j=1}^{n} a_{ij}\mathbf{z}_i(k) + \frac{\nabla f_i\big(\mathbf{x}_i(k+1)\big)}{[\mathbf{y}_i(k+1)]_i} - \frac{\nabla f_i\big(\mathbf{x}_i(k)\big)}{[\mathbf{y}_i(k)]_i},
\end{aligned}$$

where $\underline{A} = \{a_{ij}\}$ is row-stochastic. Note that the first equation is an independent algorithm, which asymptotically learns the left eigenvector, corresponding to the eigenvalue of 1, of $\underline{A}$. However, it adds nonlinearity to the overall algorithm along with additional computation and communication costs in contrast to the proposed algorithm in Eqs. (1).

Remarks: The algorithm, Eqs. (1), proposed in this letter can be viewed as related to Eq. (4) but without doubly-stochastic weights, due to which we lose the nice eigenstructure within the weight matrices. It is rather straightforward to notice that a linear extension of Eqs. (4) to the directed graphs is non-trivial as all earlier attempts were made by adding nonlinearity to the original set of equations. One of the major challenges lies in the fact that even though the contraction of a doubly-stochastic W is well-established in the subspace orthogonal to $\mathbf{1}_n$, it is not straightforward to establish simultaneous contractions for a row-stochastic matrix, $\underline{A}$, and a column-stochastic matrix, $\underline{B}$. The latter requires working with arbitrary norms (as opposed to the 2-norm applicable to doubly-stochastic matrices) and norm-equivalence constants, as we show in Lemma 1 and onwards.

III. Convergence Analysis

For the sake of analysis, we now write Eqs. (1) in matrix form. The variables $\mathbf{x}(k)$ and $\mathbf{y}(k)$ collect the local variables $\mathbf{x}_i(k)$'s and $\mathbf{y}_i(k)$'s in a vector, respectively, and $\nabla\mathbf{f}(k) = \big[\nabla f_1(\mathbf{x}_1(k))^\top, \cdots, \nabla f_n(\mathbf{x}_n(k))^\top\big]^\top$.

Let $A = \underline{A}\otimes I_p$ and $B = \underline{B}\otimes I_p$, where $\otimes$ is the Kronecker product. We now rewrite Eqs. (1) in a matrix form as follows:

$$\mathbf{x}(k+1) = A\mathbf{x}(k) - \eta\mathbf{y}(k), \tag{5a}$$

$$\mathbf{y}(k+1) = B\Big(\mathbf{y}(k) + \nabla\mathbf{f}(k+1) - \nabla\mathbf{f}(k)\Big), \tag{5b}$$

where $\mathbf{y}(0) = \nabla\mathbf{f}(0)$ and $\mathbf{x}(0)$ is arbitrary.

A. Auxiliary relations

We next start the convergence analysis with a key lemma regarding the contraction in consensus process with row- and column-stochastic weight matrices, respectively.

Lemma 1. *Consider the weight matrices, $A = \underline{A}\otimes I_p$ and $B = \underline{B}\otimes I_p$. There exist vector norms, $\|\cdot\|_A$ and $\|\cdot\|_B$, such that for all $\mathbf{a}\in\mathbb{R}^{np}$,*

$$\|A\mathbf{a} - A_\infty\mathbf{a}\|_A \le \sigma_A\|\mathbf{a} - A_\infty\mathbf{a}\|_A, \tag{6}$$

$$\|B\mathbf{a} - B_\infty\mathbf{a}\|_B \le \sigma_B\|\mathbf{a} - B_\infty\mathbf{a}\|_B, \tag{7}$$

where $0<\sigma_A<1$ and $0<\sigma_B<1$ are some constants.

Proof. Since $\underline{A}$ is irreducible, row-stochastic with positive diagonals, from Perron-Frobenius theorem we have that $\rho(\underline{A}) = 1$, every eigenvalue of $\underline{A}$ other than 1 is strictly less than $\rho(\underline{A})$, and $\boldsymbol{\pi}_r^\top$ is a strictly positive left eigenvector corresponding to the eigenvalue of 1 with $\mathbf{1}_n^\top\boldsymbol{\pi}_r = 1$; thus $\lim_{k\to\infty}\underline{A}^k = \mathbf{1}_n\boldsymbol{\pi}_r^\top$. We further have

$$A_\infty = \lim_{k\to\infty}A^k = \left(\lim_{k\to\infty}\underline{A}^k\right)\otimes I_p = \big(\mathbf{1}_n\boldsymbol{\pi}_r^\top\big)\otimes I_p.$$

It follows that

$$\begin{aligned}
AA_\infty &= (\underline{A}\otimes I_p)\Big((\mathbf{1}_n\boldsymbol{\pi}_r^\top)\otimes I_p\Big) = A_\infty,\\
A_\infty A_\infty &= \Big((\mathbf{1}_n\boldsymbol{\pi}_r^\top)\otimes I_p\Big)\Big((\mathbf{1}_n\boldsymbol{\pi}_r^\top)\otimes I_p\Big) = A_\infty.
\end{aligned}$$

Thus $AA_\infty - A_\infty A_\infty$ is a zero matrix, which leads to the following relation:

$$A\mathbf{a} - A_\infty\mathbf{a} = (A - A_\infty)(\mathbf{a} - A_\infty\mathbf{a}). \tag{8}$$

Since $\rho(A - A_\infty) = \rho((\underline{A} - \mathbf{1}_n\boldsymbol{\pi}_r^\top)\otimes I_p) < 1$, we have from Lemma 5.6.10 in [27] that there exists a matrix norm, say $|||\cdot|||_A$, such that

$$\sigma_A \triangleq |||A - A_\infty|||_A < 1.$$

Moreover, from Theorem 5.7.13 in [27], we know that for any matrix norm, $|||\cdot|||_A$, there exists a compatible vector norm, say $\|\cdot\|_A$, such that $\|X\mathbf{x}\|_A \le |||X|||_A\|\mathbf{x}\|_A$, for all matrices, X, and all vectors, $\mathbf{x}$; hence, Eq. (8) leads to

$$\begin{aligned}
\|A\mathbf{a} - A_\infty\mathbf{a}\|_A &= \|(A - A_\infty)(\mathbf{a} - A_\infty\mathbf{a})\|_A,\\
&\le |||A - A_\infty|||_A\|\mathbf{a} - A_\infty\mathbf{a}\|_A,\\
&= \sigma_A\|\mathbf{a} - A_\infty\mathbf{a}\|_A,
\end{aligned}$$

and Eq. (6) follows. Similarly, Eq. (7) follows for some matrix norm, $|||\cdot|||_B$, with $\sigma_B \triangleq |||B - B_\infty|||_B$. □

The following lemma is a direct consequence of the column-stochasticity of $\underline{B}$ and that $\mathbf{y}(0) = \nabla\mathbf{f}(0)$.

Lemma 2. *We have $(\mathbf{1}_n^\top\otimes I_p)\mathbf{y}(k) = (\mathbf{1}_n^\top\otimes I_p)\nabla\mathbf{f}(k), \forall k$.*

Proof. Recall Eq. (5b) and multiply both sides of Eq. (5b) with $\mathbf{1}_n^\top\otimes I_p$. We get

$$\begin{aligned}
&(\mathbf{1}_n^\top\otimes I_p)\mathbf{y}(k+1)\\
&= (\mathbf{1}_n^\top\otimes I_p)(\underline{B}\otimes I_p)\Big(\mathbf{y}(k) + \nabla\mathbf{f}(k+1) - \nabla\mathbf{f}(k)\Big)\\
&= (\mathbf{1}_n^\top\otimes I_p)\mathbf{y}(k) + (\mathbf{1}_n^\top\otimes I_p)\big(\nabla\mathbf{f}(k+1) - \nabla\mathbf{f}(k)\big)\\
&= (\mathbf{1}_n^\top\otimes I_p)\Big(\mathbf{y}(0) - \nabla\mathbf{f}(0)\Big) + (\mathbf{1}_n^\top\otimes I_p)\nabla\mathbf{f}(k+1)\\
&= (\mathbf{1}_n^\top\otimes I_p)\nabla\mathbf{f}(k+1),
\end{aligned}$$

which completes the proof. □

Lemma 2 shows that the average of $\mathbf{y}_i(k)$'s preserves the average of local gradients. The next lemma, from [5], [28], states that the distance to the optimal minimizer shrinks by at least a fixed ratio if we perform a gradient descent step.

Lemma 3. *Let Assumption 2 hold for the objective functions, $f_i(\mathbf{x})$, in Problem P1, and let α and β be the strong-convexity and Lipschitz-continuity constants, respectively. For any $\mathbf{x} \in \mathbb{R}^p$ and $0 < \theta < \frac{2}{\beta}$, we have for $F = \frac{1}{n}\sum_i f_i$:*

$$\|\mathbf{x} - \theta\nabla F(\mathbf{x}) - \mathbf{x}^*\|_2 \leq \tau \|\mathbf{x} - \mathbf{x}^*\|_2,$$

where $\tau = \max(|1-\alpha\theta|, |1-\beta\theta|)$.

The subsequent convergence analysis is based on deriving a contraction relationship in the proposed algorithm, i.e., $\|\mathbf{x}(k+1) - A_\infty\mathbf{x}(k+1)\|_A$, $\|A_\infty\mathbf{x}(k+1) - \mathbf{1}_n\otimes\mathbf{x}^*\|_2$, and $\|\mathbf{y}(k+1) - B_\infty\mathbf{y}(k+1)\|_B$, are bounded linearly by their values in the last iteration. We capture a relationship on these objects in the next lemmas. Before we proceed, note that all vector norms on finite-dimensional vector space are equivalent, i.e., there exists finite, positive constants, c, d, h, l, g, m, such that:

$$\|\cdot\|_A \leq c\|\cdot\|_B, \quad \|\cdot\|_2 \leq h\|\cdot\|_B, \quad \|\cdot\|_2 \leq g\|\cdot\|_A,$$
$$\|\cdot\|_B \leq d\|\cdot\|_A, \quad \|\cdot\|_B \leq l\|\cdot\|_2, \quad \|\cdot\|_A \leq m\|\cdot\|_2.$$

Lemma 4. *The following inequality holds, $\forall k$:*

$$\begin{aligned}&\|\mathbf{x}(k+1) - A_\infty\mathbf{x}(k+1)\|_A\\ &\leq \sigma_A\|\mathbf{x}(k) - A_\infty\mathbf{x}(k)\|_A + \ \eta c\|\mathbf{y}(k) - B_\infty\mathbf{y}(k)\|_B\\ &\quad + \ \eta m\|\!|\, B_\infty - A_\infty \,|\!\|_A\|\mathbf{y}(k)\|_2.\end{aligned}$$

Proof. Using Eq. (5a) and Lemma. 1, we have

$$\begin{aligned}&\|\mathbf{x}(k+1) - A_\infty\mathbf{x}(k+1)\|_A\\ &= \|A\mathbf{x}(k) - \eta\mathbf{y}(k) - A_\infty\Big(A\mathbf{x}(k) - \eta\mathbf{y}(k)\Big)\|_A,\\ &\leq \sigma_A\|\mathbf{x}(k) - A_\infty\mathbf{x}(k)\|_A + \eta\|\mathbf{y}(k) - A_\infty\mathbf{y}(k)\|_A,\\ &\leq \sigma_A\|\mathbf{x}(k) - A_\infty\mathbf{x}(k)\|_A + \eta c\|\mathbf{y}(k) - B_\infty\mathbf{y}(k)\|_B\\ &\quad + \ \eta m\|\!|\, B_\infty - A_\infty \,|\!\|_A\|\mathbf{y}(k)\|_2,\end{aligned}$$

and the lemma follows. □

Next, we develop a relation for $\|A_\infty\mathbf{x}(k+1) - \mathbf{1}_n\otimes\mathbf{x}^*\|_2$.

Lemma 5. *The following holds, $\forall k$, when $\eta < \frac{2}{n\beta\boldsymbol{\pi}_r^\top\boldsymbol{\pi}_c}$:*

$$\begin{aligned}&\|A_\infty\mathbf{x}(k+1) - \mathbf{1}_n\otimes\mathbf{x}^*\|_2\\ &\leq \eta n\beta g(\boldsymbol{\pi}_r^\top\boldsymbol{\pi}_c)\|\mathbf{x}(k) - A_\infty\mathbf{x}(k)\|_A\\ &\quad + \ \lambda\|A_\infty\mathbf{x}(k) - \mathbf{1}_n\otimes\mathbf{x}^*\|_2\\ &\quad + \ \eta h\|\!|\, A \,|\!\|_2\|\mathbf{y}(k) - B_\infty\mathbf{y}(k)\|_B,\end{aligned} \tag{9}$$

where $\lambda = \max\big(|1-\alpha n(\boldsymbol{\pi}_r^\top\boldsymbol{\pi}_c)\eta|, |1-\beta n(\boldsymbol{\pi}_r^\top\boldsymbol{\pi}_c)\eta|\big)$.

Proof. With $A_\infty = (\mathbf{1}_n\boldsymbol{\pi}_r^\top)\otimes I_p = (\mathbf{1}_n\otimes I_p)(\boldsymbol{\pi}_r^\top\otimes I_p)$ and Eq. (5a), we have

$$\begin{aligned}&\|A_\infty\mathbf{x}(k+1) - \mathbf{1}_n\otimes\mathbf{x}^*\|_2\\ &= \|A_\infty\Big(A\mathbf{x}(k) - \eta\mathbf{y}(k) + B_\infty\mathbf{y}(k)(-\eta+\eta)\Big) - \mathbf{1}_n\otimes\mathbf{x}^*\|_2,\\ &\leq \|\big((\mathbf{1}_n\boldsymbol{\pi}_r^\top)\otimes I_p\big)\mathbf{x}(k) - (\mathbf{1}_n\otimes I_p)\mathbf{x}^* - \eta A_\infty B_\infty\mathbf{y}(k)\|_2\\ &\quad + \eta h\|\!|\, A \,|\!\|_2\|\mathbf{y}(k) - B_\infty\mathbf{y}(k)\|_B.\end{aligned} \tag{10}$$

Since the last term above matches with the last term in Eq. (9), what is left is to manipulate the first term. Before we proceed, define $\nabla F(k) = \nabla F\big((\boldsymbol{\pi}_r^\top\otimes I_p)\mathbf{x}(k)\big)$, which is the global gradient evaluated at $(\boldsymbol{\pi}_r^\top\otimes I_p)\mathbf{x}(k)$. Note that

$$A_\infty B_\infty = \Big((\mathbf{1}_n\boldsymbol{\pi}_r^\top)\otimes I_p\Big)\Big((\boldsymbol{\pi}_c\mathbf{1}_n^\top)\otimes I_p\Big) = \boldsymbol{\pi}_r^\top\boldsymbol{\pi}_c\Big((\mathbf{1}_n\mathbf{1}_n^\top)\otimes I_p\Big).$$

We have

$$\begin{aligned}&\|\big((\mathbf{1}_n\boldsymbol{\pi}_r^\top)\otimes I_p\big)\mathbf{x}(k) - (\mathbf{1}_n\otimes I_p)\mathbf{x}^* - \eta A_\infty B_\infty\mathbf{y}(k)\|_2\\ &\leq \Big\|(\mathbf{1}_n\otimes I_p)\Big((\boldsymbol{\pi}_r^\top\otimes I_p)\mathbf{x}(k) - \mathbf{x}^* - n\eta(\boldsymbol{\pi}_r^\top\boldsymbol{\pi}_c)\nabla F(k)\Big)\Big\|_2\\ &+ \eta(\boldsymbol{\pi}_r^\top\boldsymbol{\pi}_c)\Big\|n(\mathbf{1}_n\otimes I_p)\nabla F(k) - (\mathbf{1}_n\otimes I_p)(\mathbf{1}_n^\top\otimes I_p)\mathbf{y}(k)\Big\|_2\\ &:= s_1 + \eta s_2.\end{aligned}$$

From Lemma 3, we have that if $0 < \eta < \frac{2}{n\beta\boldsymbol{\pi}_r^\top\boldsymbol{\pi}_c}$,

$$s_1 \leq \lambda\|A_\infty\mathbf{x}(k) - \mathbf{1}_n\otimes\mathbf{x}^*\|_2,$$

where $\lambda = \max\big(|1-\alpha n(\boldsymbol{\pi}_r^\top\boldsymbol{\pi}_c)\eta|, |1-\beta n(\boldsymbol{\pi}_r^\top\boldsymbol{\pi}_c)\eta|\big)$. Recall that $(\mathbf{1}_n^\top\otimes I_p)\mathbf{y}(k) = (\mathbf{1}_n^\top\otimes I_p)\nabla\mathbf{f}(k), \forall k$, from Lemma 2 and Assumption 2, we have

$$s_2 \leq n\beta g(\boldsymbol{\pi}_r^\top\boldsymbol{\pi}_c)\|\mathbf{x}(k) - A_\infty\mathbf{x}(k)\|_A.$$

The lemma follows by using the above bounds in Eq. (10). □

Next, we develop a relation for $\|\mathbf{y}(k+1) - B_\infty\mathbf{y}(k+1)\|_B$.

Lemma 6. *The following inequality holds, $\forall k$:*

$$\begin{aligned}&\|\mathbf{y}(k+1) - B_\infty\mathbf{y}(k+1)\|_B\\ &\leq \sigma_B\beta lg\|\!|\, A - I_{np} \,|\!\|_2\|\mathbf{x}(k) - A_\infty\mathbf{x}(k)\|_A\\ &\quad + \sigma_B\|\mathbf{y}(k) - B_\infty\mathbf{y}(k)\|_B + \eta\sigma_B\beta l\|\mathbf{y}(k)\|_2.\end{aligned}$$

Proof. We note that

$$\begin{aligned}&\|\mathbf{y}(k+1) - B_\infty\mathbf{y}(k+1)\|_B\\ &= \|B\Big(\mathbf{y}(k) + \nabla\mathbf{f}(k+1) - \nabla\mathbf{f}(k)\Big)\\ &\quad - \ B_\infty B\Big(\mathbf{y}(k) + \nabla\mathbf{f}(k+1) - \nabla\mathbf{f}(k)\Big)\|_B,\\ &\leq \sigma_B\|\mathbf{y}(k) - B_\infty\mathbf{y}(k)\|_B + \sigma_B\beta l\|\mathbf{x}(k+1) - \mathbf{x}(k)\|_2,\end{aligned} \tag{11}$$

because of Lemma 1. Now we analyze $\|\mathbf{x}(k+1) - \mathbf{x}(k)\|_2$.

$$\begin{aligned}&\|\mathbf{x}(k+1) - \mathbf{x}(k)\|_2\\ &= \|A\mathbf{x}(k) - \eta\mathbf{y}(k) - \mathbf{x}(k)\|_2,\\ &\leq \|(A - I_{np})\big(\mathbf{x}(k) - A_\infty\mathbf{x}(k)\big)\|_2 + \eta\|\mathbf{y}(k)\|_2,\\ &\leq \|\!|\, A - I_{np} \,|\!\|_2 g\|\mathbf{x}(k) - A_\infty\mathbf{x}(k)\|_A + \eta\|\mathbf{y}(k)\|_2.\end{aligned} \tag{12}$$

The lemma follows by plugging Eq. (12) into Eq. (11). □

The last step to complete the contraction relationship is to bound $\|\mathbf{y}(k)\|_2$ in terms of $\|\mathbf{x}(k) - A_\infty\mathbf{x}(k)\|_A$, $\|A_\infty\mathbf{x}(k) - \mathbf{1}_n\otimes\mathbf{x}^*\|_2$, and $\|\mathbf{y}(k) - B_\infty\mathbf{y}(k)\|_B$.

Lemma 7. *The following inequality holds, $\forall k$:*

$$\begin{aligned}\|\mathbf{y}(k)\|_2 \leq\ & g\beta\|\!|\, B_\infty \,|\!\|_2\|\mathbf{x}(k) - A_\infty\mathbf{x}(k)\|_A\\ &+ \ \beta\|\!|\, B_\infty \,|\!\|_2\|A_\infty\mathbf{x}(k) - \mathbf{1}_n\otimes\mathbf{x}^*\|_2\\ &+ \ h\|\mathbf{y}(k) - B_\infty\mathbf{y}(k)\|_B.\end{aligned}$$

Proof. Recall that $B_\infty = (\boldsymbol{\pi}_c\otimes I_p)(\mathbf{1}_n^\top\otimes I_p)$. We have

$$\|\mathbf{y}(k)\|_2 \leq h\|\mathbf{y}(k) - B_\infty\mathbf{y}(k)\|_B + \|B_\infty\mathbf{y}(k)\|_2. \tag{13}$$

We next bound $\|B_\infty \mathbf{y}(k)\|_2$:

$$\begin{aligned}
\|B_\infty \mathbf{y}(k)\|_2 &= \|(\boldsymbol{\pi}_c \otimes I_p)(\mathbf{1}_n^\top \otimes I_p)\mathbf{y}(k)\|_2 \\
&= \|\boldsymbol{\pi}_c\|_2 \|(\mathbf{1}_n^\top \otimes I_p)\nabla \mathbf{f}(k)\|_2 \\
&= \|\boldsymbol{\pi}_c\|_2 \left\| \sum_{i=1}^n \nabla f_i(\mathbf{x}_i(k)) - \sum_{i=1}^n \nabla f_i(\mathbf{x}^*) \right\|_2 \\
&\leq \|\boldsymbol{\pi}_c\|_2 \beta \sum_{i=1}^n \|\mathbf{x}_i(k) - \mathbf{x}^*\|_2 \\
&\leq \|\boldsymbol{\pi}_c\|_2 \beta \sqrt{n} \|\mathbf{x}(k) - \mathbf{1}_n \otimes \mathbf{x}^*\|_2, \\
&\leq |\!|\!| B_\infty |\!|\!|_2 \beta g \|\mathbf{x}(k) - A_\infty \mathbf{x}(k)\|_A \\
&\quad + |\!|\!| B_\infty |\!|\!|_2 \beta \|A_\infty \mathbf{x}(k) - \mathbf{1}_n \otimes \mathbf{x}^*\|_2, \quad (14)
\end{aligned}$$

where the second last inequality uses Jensen's inequality and the last inequality uses the fact that $|\!|\!| B_\infty |\!|\!|_2 = \sqrt{n}\|\boldsymbol{\pi}_c\|_2$. The lemma follows by plugging Eqs. (14) into Eq. (13). □

Before the main result, we present an additional lemma from matrix perturbation theory.

Lemma 8. *(Theorem 6.3.12 in [27]) Let $X, E \in \mathbb{R}^{n\times n}$ and let q be a simple eigenvalue of X. Let $\mathbf{v}$ and $\mathbf{w}$ be, respectively, the right and left eigenvectors of X corresponding to the eigenvalue q. Then,*

(i) for each $\epsilon > 0$, there exists a $\delta > 0$ such that, $\forall t \in \mathbb{C}$ with $|t| < \delta$, there is a unique eigenvalue $q(t)$ of $X + tE$ such that $\left| q(t) - q - t\frac{\mathbf{w}^H E \mathbf{v}}{\mathbf{w}^H \mathbf{v}} \right| \leq |t|\epsilon$,

(ii) $q(t)$ is continuous at $t = 0$, and $\lim_{t\to 0} q(t) = q$,

(iii) $q(t)$ is differentiable at $t = 0$, $\frac{dq(t)}{dt}|_{t=0} = \frac{\mathbf{w}^H E \mathbf{v}}{\mathbf{w}^H \mathbf{v}}$.

B. Main results

With the help of the auxiliary relations developed in the previous subsection, we now present the main result.

Theorem 1. *If $0 < \eta < \frac{2}{n\beta \boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c}$, we have the following linear matrix inequality (entry-wise):*

$$\mathbf{t}(k+1) \leq J(\eta)\mathbf{t}(k), \ \forall k, \quad (15)$$

where $\mathbf{t}(k) \in \mathbb{R}^3$ and $J(\eta) \in \mathbb{R}^{3\times 3}$ are defined as follows:

$$\mathbf{t}(k) = \begin{bmatrix} \|\mathbf{x}(k) - A_\infty \mathbf{x}(k)\|_A \\ \|A_\infty \mathbf{x}(k) - \mathbf{1}_n \otimes \mathbf{x}^*\|_2 \\ \|\mathbf{y}(k) - B_\infty \mathbf{y}(k)\|_B \end{bmatrix}, \quad (16)$$

$$J(\eta) = \begin{bmatrix} \sigma_A + a_1\eta & a_2\eta & a_3\eta \\ a_4\eta & \lambda & a_5\eta \\ a_6 + a_7\eta & a_8\eta & \sigma_B + a_9\eta \end{bmatrix}, \quad (17)$$

with the constants a_i's being

$$\begin{aligned}
a_1 &= mg\beta |\!|\!| B_\infty |\!|\!|_2 |\!|\!| A_\infty - B_\infty |\!|\!|_A, \\
a_2 &= m\beta |\!|\!| B_\infty |\!|\!|_2 |\!|\!| A_\infty - B_\infty |\!|\!|_A, \\
a_3 &= c + mh |\!|\!| A_\infty - B_\infty |\!|\!|_A, \\
a_4 &= n\beta g(\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c), \\
a_5 &= h |\!|\!| A_\infty |\!|\!|_2, \\
a_6 &= g\sigma_B l\beta |\!|\!| A - I_{np} |\!|\!|_2, \\
a_7 &= g\sigma_B l\beta^2 |\!|\!| B_\infty |\!|\!|_2, \\
a_8 &= \sigma_B l\beta^2 |\!|\!| B_\infty |\!|\!|_2, \\
a_9 &= h\sigma_B l\beta.
\end{aligned}$$

When the step-size η is sufficiently small, the spectral radius of $J(\eta)$, $\rho(J(\eta))$, is strictly less than 1, and therefore $\|\mathbf{x}(k) - \mathbf{1}_n \otimes \mathbf{x}^*\|_2$ converges to zero geometrically at the rate of $O(\rho(J(\eta))^k)$.

Proof. Combining the results of Lemmas 4–7, one can verify Eq. (15). Next, we show that when the step-size η is sufficiently small, the spectral radius of $J(\eta)$ is strictly less than 1. Recall from Lemma 5 that $\lambda = \max\left(\left|1 - \alpha n(\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c)\eta\right|, \left|1 - \beta n(\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c)\eta\right|\right)$. Therefore, when $\eta < \frac{1}{n\beta(\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c)}$, $\lambda = 1 - \alpha n(\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c)\eta$, since $\alpha \leq \beta$; see, e.g., [28] for details. We now split the matrix, $J(\eta)$, into the sum of a fixed matrix and another perturbation matrix as a function η:

$$\begin{aligned}
J(\eta) &= \begin{bmatrix} \sigma_A & 0 & 0 \\ 0 & 1 & 0 \\ a_6 & 0 & \sigma_B \end{bmatrix} + \eta \begin{bmatrix} a_1 & a_2 & a_3 \\ a_4 & -\alpha n(\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c) & a_5 \\ a_7 & a_8 & a_9 \end{bmatrix} \\
&:= J_0 + \eta E.
\end{aligned}$$

Clearly, the spectral radius of J_0 is 1 since both σ_A and σ_B are in $(0, 1)$. It is straightforward to verify that the right and left eigenvector corresponding to the eigenvalue of 1 of J_0 is $\mathbf{v} = [0, 1, 0]^\top$. Denote by $q(\eta)$, the eigenvalues of $J(\eta)$ as a function of η. From Lemma 8, since 1 is a simple eigenvalue of $J(0)$,

$$\left.\frac{dq(\eta)}{d\eta}\right|_{\eta=0, q=1} = \frac{\mathbf{v}^\top E \mathbf{v}}{\mathbf{v}^\top \mathbf{v}} = -\alpha n(\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c),$$

i.e., $\frac{dq}{d\eta}|_{\eta=0,q=1} < 0$ and the spectral radius of $J(\eta)$ is strictly less than 1 as η slightly increases from zero. This is because the eigenvalues are continuous functions of the elements of the matrix, and the theorem follows. □

Remarks: Note that the notion of a sufficiently small step-size is not uncommon in the literature, see, e.g., [4]–[6], [22]–[25]. Typically, an applicable upper bound on the step-size is a function of all of the network and algorithm parameters and thus is practically infeasible to compute in distributed settings; in addition, several constants, e.g., the norm equivalence constants, and network weights are arbitrary, see [22]–[25] for further details.

IV. Numerical Experiments

We consider a binary classification problem in the distributed setting, where we use logistic loss function to train a linear classifier. Each agent i has access to m_i training data, $(\mathbf{c}_{ij}, y_{ij}) \in \mathbb{R}^p \times \{-1, +1\}$, where $\mathbf{c}_{ij}$ contains p features of the jth training data at agent i and y_{ij} is the corresponding binary label. For privacy issues, agents do not share training data with each other. In order to use the entire data set for training, the network of agents cooperatively solves the following distributed logistic regression problem:

$$\min_{\mathbf{w},b} F(\mathbf{w}, b) = \sum_{i=1}^n \sum_{j=1}^{m_i} \ln\left[1 + e^{-(\mathbf{w}^\top \mathbf{c}_{ij} + b)y_{ij}}\right] + \frac{\xi}{2}\|\mathbf{w}\|_2^2,$$

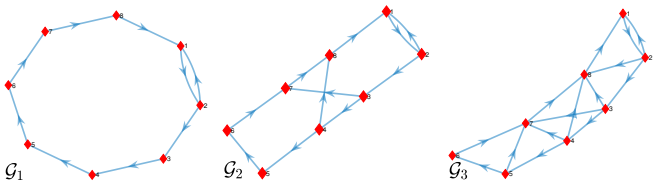

Fig. 1. Strongly-connected but unbalanced directed graphs.

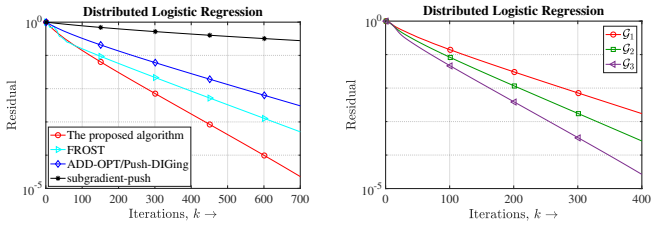

Fig. 2. (Left) Comparison across different algorithms. (Right) Proposed algorithm over different graphs.

where the private function at each agent, i, is given by:

$$f_i(\mathbf{w}, b) = \sum_{j=1}^{m_i} \ln\left[1 + e^{-(\mathbf{w}^\top \mathbf{c}_{ij} + b) y_{ij}}\right] + \frac{\xi}{2n}\|\mathbf{w}\|_2^2.$$

In our setting, $n = 8$, $p = 5$. The feature vectors, $\mathbf{c}_{ij}$'s, are Gaussian with zero mean and variance 2. The binary labels are randomly generated from standard Bernoulli distribution. We first compare the performance of the proposed algorithm in this paper, with ADD-OPT/Push-DIGing [22], [23], FROST [25], and subgradient-push [15], [16], over the leftmost directed graph, $\mathcal{G}_1$, shown in Fig. 1. The simulation results are shown in the left figure in Fig. 2. Next, we evaluate the proposed algorithm on the three different directed graphs, $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3$, shown in Fig. 1, where each graph to the right has a few more edges compared to the one on its left. The simulation results are shown in the right figure in Fig. 2. In both cases, we plot the average of the residuals at each agent, $\frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x}_i(k) - \mathbf{x}^*\|_2$. We note that the proposed linear algorithm achieves a geometric (linear on the log-scale) convergence speed comparable to other fast algorithms over directed graphs but with less computation and communication. These simulations confirm the theoretical findings in this letter.

V. Conclusions

In this letter, we describe a linear distributed algorithm for optimization over directed graphs that can be seen as a generalization of earlier work over undirected graphs. Under the assumptions that the objective functions are strongly-convex and have Lipschitz-continuous gradients, the proposed algorithm achieves a geometric convergence to the global optimal. Our analysis is based on a novel approach where we establish simultaneous contractions of both row- and column-stochastic matrices under some arbitrary norms. We then use an elegant result from matrix perturbation theory to develop the conditions for convergence.

References

[1] A. Nedić and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Trans. on Automatic Control*, vol. 54, no. 1, pp. 48–61, Jan. 2009.

[2] K. Yuan, Q. Ling, and W. Yin, "On the convergence of decentralized gradient descent," *SIAM Journal on Optimization*, vol. 26, no. 3, pp. 1835–1854, Sep. 2016.

[3] W. Shi, Q. Ling, G. Wu, and W Yin, "Extra: An exact first-order algorithm for decentralized consensus optimization," *SIAM Journal on Optimization*, vol. 25, no. 2, pp. 944–966, 2015.

[4] J. Xu, Sj Zhu, Yj Cj Soh, and L. Xie, "Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes," in *IEEE 54th Annual Conference on Decision and Control*, 2015, pp. 2055–2060.

[5] G. Qu and N. Li, "Harnessing smoothness to accelerate distributed optimization," *IEEE Trans. on Control of Network Systems*, Apr. 2017.

[6] G. Qu and N. Li, "Accelerated distributed Nesterov gradient descent," *Arxiv: https://arxiv.org/abs/1705.07176*, May 2017.

[7] M. Zhu and S. Martínez, "Discrete-time dynamic average consensus," *Automatica*, vol. 46, no. 2, pp. 322–329, 2010.

[8] J. F. C. Mota, J. M. F. Xavier, P. M. Q. Aguiar, and M. Püschel, "Distributed basis pursuit," *IEEE Transactions on Signal Processing*, vol. 60, no. 4, pp. 1942–1956, Apr. 2012.

[9] D. Jakovetic, "A unification, generalization, and acceleration of exact distributed first order methods," *arXiv preprint arXiv:1709.01317*, 2017.

[10] H. Raja and W. U. Bajwa, "Cloud K-SVD: A collaborative dictionary learing algorithm for big, distributed data," *IEEE Trans. Signal Processing*, vol. 64, no. 1, pp. 173–188, Jan. 2016.

[11] S. Lee and M. M. Zavlanos, "Approximate projection methods for decentralized optimization with functional constraints," *IEEE Transactions on Automatic Control*, 2017.

[12] F. Mansoori and E. Wei, "Superlinearly convergent asynchronous distributed network Newton method," in *56th IEEE Annual Conference on Decision and Control*, Dec. 2017, pp. 2874–2879.

[13] B. Ying and A. H. Sayed, "Performance limits of stochastic sub-gradient learning, part II: Multi-agent case," *Signal Processing*, vol. 144, pp. 253–264, Mar. 2018.

[14] B. Gharesifard and J. Cortés, "Distributed strategies for generating weight-balanced and doubly stochastic digraphs," *European Journal of Control*, vol. 18, no. 6, pp. 539–557, 2012.

[15] K. I. Tsianos, S. Lawlor, and M. G. Rabbat, "Push-sum distributed dual averaging for convex optimization," in *51st IEEE Annual Conference on Decision and Control*, Maui, Hawaii, Dec. 2012, pp. 5453–5458.

[16] A. Nedić and A. Olshevsky, "Distributed optimization over time-varying directed graphs," *IEEE Trans. on Automatic Control*, vol. 60, no. 3, pp. 601–615, Mar. 2015.

[17] D. Kempe, A. Dobra, and J. Gehrke, "Gossip-based computation of aggregate information," in *44th Annual IEEE Symposium on Foundations of Computer Science*, Oct. 2003, pp. 482–491.

[18] C. Xi, Q. Wu, and U. A. Khan, "On the distributed optimization over directed networks," *Neurocomputing*, vol. 267, pp. 508–515, Dec. 2017.

[19] C. Xi and U. A. Khan, "Distributed subgradient projection algorithm over directed graphs," *IEEE Trans. on Automatic Control*, vol. 62, no. 8, pp. 3986–3992, Oct. 2016.

[20] K. Cai and H. Ishii, "Average consensus on general strongly connected digraphs," *Automatica*, vol. 48, no. 11, pp. 2750 – 2761, 2012.

[21] C. Xi and U. A. Khan, "DEXTRA: A fast algorithm for optimization over directed graphs," *IEEE Trans. on Automatic Control*, vol. 62, no. 10, pp. 4980–4993, Oct. 2017.

[22] C. Xi, R. Xin, and U. A. Khan, "ADD-OPT: Accelerated distributed directed optimization," *IEEE Trans. on Automatic Control*, Aug. 2017, *in press*.

[23] A. Nedić, A. Olshevsky, and W. Shi, "Achieving geometric convergence for distributed optimization over time-varying graphs," *SIAM Journal of Optimization*, Dec. 2017.

[24] C. Xi, V. S. Mai, R. Xin, E. Abed, and U. A. Khan, "Linear convergence in optimization over directed graphs with row-stochastic matrices," *IEEE Trans. on Automatic Control*, Jan. 2018, *in press*.

[25] R. Xin, C. Xi, and U. A. Khan, "FROST – Fast row-stochastic optimization with uncoordinated step-sizes," *Arxiv: https://arxiv.org/abs/1803.09169*, Mar. 24th 2018.

[26] F. Bullo, J. Cortes, and S. Martinez, *Distributed Control of Robotic Networks*, Princeton University: Applied Mathematics Series, 2009.

[27] R. A. Horn and C. R. Johnson, *Matrix Analysis, 2nd ed.*, Cambridge University Press, New York, NY, 2013.

[28] S. Bubeck, "Convex optimization: Algorithms and complexity," *arXiv preprint arXiv:1405.4980*, 2014.