ADD-OPT: Accelerated Distributed Directed Optimization

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

Abstract—In this paper, we consider distributed optimization problems where the goal is to minimize a sum of objective functions over a multi-agent network. We focus on the case when the inter-agent communication is described by a strongly-connected, *directed* graph. The proposed algorithm, ADD-OPT (Accelerated Distributed Directed Optimization), achieves the best known convergence rate for this class of problems, $O(\mu^k), 0 < \mu < 1$, given strongly-convex, objective functions with globally Lipschitz-continuous gradients, where k is the number of iterations. Moreover, ADD-OPT supports a wider and more realistic range of step-sizes in contrast to existing work. In particular, we show that ADD-OPT converges for arbitrarily small (positive) step-sizes. Simulations further illustrate our results.

Index Terms—Distributed optimization, directed graph, linear convergence, DEXTRA.

I. INTRODUCTION

In this paper, we consider distributed optimization problems where the goal is to minimize a sum of objective functions over a multi-agent network. Formally, we consider a decision variable, $\mathbf{z} \in \mathbb{R}^p$, and a strongly-connected network containing n agents, where each agent, i, only has access to a local objective function, $f_i : \mathbb{R}^p \to \mathbb{R}$. The goal is to have each agent minimize the sum of objectives, $\sum_{i=1}^n f_i(\mathbf{z})$, via information exchange with the neighbors. This formulation has gained great interest due to its widespread applications in, e.g., large-scale machine learning, [1, 2], model-predictive control, [3], cognitive networks, [4, 5], source localization, [6, 7], resource scheduling, [8], and message routing, [9].

Most of the existing algorithms assume information exchange over undirected networks (graphs), where the communication between the agents is bidirectional, i.e., if agent i sends information to agent j then agent j can also send information to agent i. Related work includes Distributed Gradient Descent (DGD), [10–13], which achieves $O(\frac{\ln k}{\sqrt{k}})$ convergence for arbitrary convex functions, and $O(\frac{\ln k}{k})$ for strongly-convex functions, where k is the number of iterations. The convergence rates can be accelerated with an additional Lipschitz-continuity assumption on the associated gradient. For example, see DGD [14] that converges at $O(\frac{1}{k})$ for general convex functions but within a ball around the optimal solution, whereas, it converges linearly to the optimal solution for strongly-convex functions. The distributed Nesterov's method, [15], converges at $O(\frac{\ln k}{k^2})$ for general convex functions. Of significant relevance is EXTRA, [16], which converges to the optimal solution at $O(\frac{1}{k})$ for general convex functions and is linear for strongly-convex functions. The work in [17] improves EXTRA by relaxing the weight matrices

to be asymmetric. Besides the gradient-based methods, the distributed implementation of ADMM, [18–20], has also been considered over undirected graphs.

The aforementioned methods, [10–20], are applicable to undirected graphs that allow the use of doubly-stochastic weight matrices; row-stochasticity guarantees that all agents reach consensus, while the column-stochasticity ensures that each local gradient contributes equally to the global objective, [21]. On the contrary, when the underlying graph is directed, the weight matrix may only be row-stochastic or column-stochastic but not both. In this paper, we provide a distributed optimization algorithm that does not require doubly-stochastic weights and thus is applicable to directed graphs (digraphs). See [22, 23] for work on balancing the weights in strongly-connected digraphs.

Optimization in continuous-time over weight-balanced digraphs has been studied earlier in [24, 25]. Existing discrete-time algorithms include the following: Gradient-Push (GP), [26–29], that combines DGD, [10], and push-sum consensus, [30, 31]; Directed-Distributed Gradient Descent (D-DGD), [21, 32], which uses Cai and Ishii's work on surplus consensus, [33], and combines it with DGD; and [34], where the authors apply the weight-balancing technique, [35], to DGD. These gradient-based methods, [21, 26–29, 32, 34], restricted by the diminishing step-size, converge relatively slowly at $O(\frac{\ln k}{\sqrt{k}})$. When the objective functions are strongly-convex, the convergence rate can be accelerated to $O(\frac{\ln k}{k})$, [36].

A recent paper proposed a fast distributed algorithm, termed DEXTRA, [37, 38], to solve the distributed consensus optimization problem over directed graphs. By combining the push-sum protocol, [30, 31], and EXTRA, [16], DEXTRA achieves a linear convergence rate given that the objective functions are strongly-convex. However, a limitation of DEXTRA is a restrictive step-size range, i.e., the greatest lower bound of DEXTRA's step-size is strictly greater than zero. In particular, DEXTRA requires the step-size, α, to follow $\alpha \in (\underline{\alpha}, \overline{\alpha})$, where $\underline{\alpha} > 0$. Estimating $\underline{\alpha}$ in a distributed setting is challenging because it may require global knowledge. In contrast if $\underline{\alpha} = 0$, agents can pick a small enough positive constant to ensure the convergence. In this paper, we propose ADD-OPT (Accelerated Distributed Directed Optimization) to address the step-size limitation inherent to DEXTRA. In particular, ADD-OPT's step-size follows $\alpha \in (0, \overline{\alpha})$, i.e., $\underline{\alpha} = 0$, ensuring that the lower bound of ADD-OPT's step-size does not require any global knowledge. We show that ADD-OPT converges linearly for strongly-convex functions.

The remainder of the paper is organized as follows. Section II formulates the problem and describes ADD-OPT. We also present appropriate assumptions in Section II. Section III

states the main convergence results. In Section IV, we present some lemmas as the basis of the proof of ADD-OPT's convergence. The proof of main results is provided in Section V. We show numerical results in Section VI and Section VII contains the concluding remarks.

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; $\mathbf{1}_n$ and $\mathbf{0}_n$ are the n-dimensional column vectors of all 1's and 0's, respectively. We denote by $A \otimes B$, the Kronecker product of two matrices, A and B. For any $f(\mathbf{x})$, $\nabla f(\mathbf{x})$ denotes the gradient of f at $\mathbf{x}$. The spectral radius of a matrix, A, is represented by $\rho(A)$. For an irreducible, column-stochastic matrix, A, we denote its right and left eigenvectors corresponding to the eigenvalue of 1 by $\boldsymbol{\pi}$ and $\mathbf{1}_n^\top$, respectively, such that $\mathbf{1}_n^\top \boldsymbol{\pi} = 1$. Depending on its argument, we denote $\| \cdot \|$ as either a particular matrix norm, the choice of which will be clear in Lemma 2, or a vector norm that is compatible with this particular matrix norm, i.e., $\|A\mathbf{x}\| \leq \|A\|\|\mathbf{x}\|$ for all matrices, A, and all vectors, $\mathbf{x}$. The notation $\| \cdot \|_2$ denotes the Euclidean norm of vectors and matrices. Since all vector norms on finite-dimensional vector space are equivalent, we have the following: $c'\| \cdot \| \leq \| \cdot \|_2 \leq c\| \cdot \|, d'\| \cdot \|_2 \leq \| \cdot \| \leq d\| \cdot \|_2$, where c', c, d', d are some positive constants.

II. ADD-OPT DEVELOPMENT

In this section, we formulate the optimization problem and describe ADD-OPT. We first derive an informal but intuitive proof showing that ADD-OPT enables the agents to achieve consensus and reach the optimal solution of Problem P1, described below. After propose ADD-OPT, we relate it to DEXTRA and discuss the applicable range of step-sizes. Formal convergence results are deferred to Sections III.

Consider a strongly-connected network of n agents communicating over a directed graph, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the 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, $j \rightarrow i$. Define $\mathcal{N}_i^{\text{in}}$ to be 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. Note that in a directed graph when $(i, j) \in \mathcal{E}$, it is not necessary that $(j, i) \in \mathcal{E}$. Consequently, $\mathcal{N}_i^{\text{in}} \neq \mathcal{N}_i^{\text{out}}$, in general. We assume that each agent i knows[1] its out-degree (the number of out-neighbors), denoted by $|\mathcal{N}_i^{\text{out}}|$; see [39] for details. We focus on solving a convex optimization problem that is distributed over the above multi-agent network. In particular, the network of agents cooperatively solves the following optimization problem:

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

where each local objective function, $f_i : \mathbb{R}^p \rightarrow \mathbb{R}$ is known only by agent i. We assume that each local function, $f_i(z)$, is strongly-convex and differentiable, whereas the optimal solution of Problem P1 exists and is finite. Our goal is to develop

[1] Such an assumption is standard in the related literature, see e.g., [21, 26–29, 32, 34, 37].

a distributed algorithm such that each agent converges to the global solution of Problem P1 via exchanging information with nearby agents over a directed graph. We formalize the set of assumptions as follows. These assumptions are standard in the literature for optimization of smooth convex functions, see e.g., [14, 16, 37].

Assumption A1. *The communication graph, $\mathcal{G}$, is a strongly-connected digraph. Each agent in the network has the knowledge of its out-degree.*

Assumption A2 (Lipschitz-continuous gradients and strong–convexity). *Each local function, f_i, is differentiable and strongly-convex, and the gradient is globally Lipschitz-continuous, i.e., for any i and $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^p$,*

(a) there exists a positive constant l such that

$$\|\nabla f_i(\mathbf{z}_1) - \nabla f_i(\mathbf{z}_2)\|_2 \leq l \|\mathbf{z}_1 - \mathbf{z}_2\|_2;$$

(b) there exists a positive constant s such that,

$$f_i(\mathbf{z}_1) - f_i(\mathbf{z}_2) \leq \nabla f_i(\mathbf{z}_1)^\top (\mathbf{z}_1 - \mathbf{z}_2) - \frac{s}{2} \|\mathbf{z}_1 - \mathbf{z}_2\|_2^2.$$

Clearly, the Lipschitz-continuity and strongly-convexity constants for the global objective function $f(\mathbf{z})$ are nl and ns, respectively.

Assumption A3. *The optimal solution exists and is bounded and unique. In particular, we denote $\underline{\mathbf{z}}^* \in \mathbb{R}^p$ the optimal solution, i.e.,*

$$\underline{\mathbf{z}}^* = \min_{\mathbf{z} \in \mathbb{R}^p} f(\mathbf{z}).$$

A. ADD-OPT Algorithm

To solve Problem P1, we describe the implementation of ADD-OPT as follows. Each agent, $j \in \mathcal{V}$, maintains three vector variables: $\mathbf{x}_k^j, \mathbf{z}_k^j, \mathbf{w}_k^j$, all in $\mathbb{R}^p$, as well as a scalar variable, $\mathsf{y}_k^j \in \mathbb{R}$, where k is the discrete-time index. At the kth iteration, agent j assigns a weight to its states: $a_{ij}\mathbf{x}_k^j$, $a_{ij}\mathbf{w}_k^j$, and $a_{ij}\mathsf{y}_k^j$; and sends these to each of its out-neighbors, $i \in \mathcal{N}_j^{\text{out}}$, where the weights, a_{ij}'s are such that:

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

With agent i receiving the information from its in-neighbors, it updates $\mathbf{x}_{k+1}^i$, y_{k+1}^i, $\mathbf{z}_{k+1}^i$ and $\mathbf{w}_{k+1}^i$ as follows:

$$\mathbf{x}_{k+1}^i = \sum_{j \in \mathcal{N}_i^{\text{in}}} a_{ij} \mathbf{x}_k^j - \alpha \mathbf{w}_k^i, \quad (2a)$$

$$\mathsf{y}_{k+1}^i = \sum_{j \in \mathcal{N}_i^{\text{in}}} a_{ij} \mathsf{y}_k^j, \quad (2b)$$

$$\mathbf{z}_{k+1}^i = \frac{\mathbf{x}_{k+1}^i}{\mathsf{y}_{k+1}^i}, \quad (2c)$$

$$\mathbf{w}_{k+1}^i = \sum_{j \in \mathcal{N}_i^{\text{in}}} a_{ij} \mathbf{w}_k^j + \nabla f_i(\mathbf{z}_{k+1}^i) - \nabla f_i(\mathbf{z}_k^i). \quad (2d)$$

In the above, $\nabla f_i(\mathbf{z}_k^i)$ is the gradient of $f_i(\mathbf{z})$ at $\mathbf{z} = \mathbf{z}_k^i$. The step-size, α, is a positive number within a certain interval. We will explicitly show the range of α in Section III. For any

agent i, it is initialized with arbitrary vectors, $\mathbf{x}_0^i$ and $\mathbf{z}_0^i$, $\mathbf{w}_0^i = \nabla f_i(\mathbf{z}_0^i)$, and $\mathbf{y}_0^i = 1$. It is worth noting that $\mathbf{y}_k^i \neq 0$, $\forall k$, given its initial condition and Assumption A1, [40]. We note that Eq. (1) leads to a column-stochastic weight matrix, $\underline{A} = \{a_{ij}\}$, by only requiring each agent to know its out-degree. It is indeed possible to construct such weights, e.g., by choosing

$$a_{ij} = \begin{cases} 1/|\mathcal{N}_j^{\text{out}}|, & i \in \mathcal{N}_j^{\text{out}}, \\ 0, & \text{otherwise}, \end{cases} , \forall j. \qquad (3)$$

For analysis purposes, we now write Eq. (2) in a matrix form. We use the following notation:

$$\mathbf{x}_k = \begin{bmatrix} \mathbf{x}_k^1 \\ \vdots \\ \mathbf{x}_k^n \end{bmatrix}, \ \mathbf{w}_k = \begin{bmatrix} \mathbf{w}_k^1 \\ \vdots \\ \mathbf{w}_k^n \end{bmatrix}, \ \mathbf{z}_k = \begin{bmatrix} \mathbf{z}_k^1 \\ \vdots \\ \mathbf{z}_k^n \end{bmatrix}, \qquad (4)$$

$$\nabla \mathbf{f}_k = \begin{bmatrix} \nabla \mathbf{f}_1(\mathbf{z}_k^1) \\ \vdots \\ \nabla \mathbf{f}_n(\mathbf{z}_k^n) \end{bmatrix}, \ \mathbf{y}_k = \begin{bmatrix} y_k^1 \\ \vdots \\ y_k^n \end{bmatrix}. \qquad (5)$$

Let $\underline{A} \in \mathbb{R}^{n \times n}$ be the weighted adjacency matrix, i.e., the collection of weights, a_{ij}; define

$$A = \underline{A} \otimes I_p, \qquad (6)$$
$$Y_k = \operatorname{diag}(\mathbf{y}_k) \otimes I_p. \qquad (7)$$

where '$\otimes$' is the Kronecker product. Clearly, we have $A, Y_k \in \mathbb{R}^{np \times np}$, and A is a column-stochastic matrix. Given that $\mathbf{y}_0 = \mathbf{1}_n$, the graph, $\mathcal{G}$, is strongly-connected and the corresponding weight matrix, $\underline{A}$, is non-negative, Y_k is invertible for any k, [40]. Then, we can write Eq. (2) in the matrix form, equivalently, as follows:

$$\mathbf{x}_{k+1} = A\mathbf{x}_k - \alpha \mathbf{w}_k, \qquad (8a)$$
$$\mathbf{y}_{k+1} = \underline{A}\mathbf{y}_k, \qquad (8b)$$
$$\mathbf{z}_{k+1} = Y_{k+1}^{-1}\mathbf{x}_{k+1}, \qquad (8c)$$
$$\mathbf{w}_{k+1} = A\mathbf{w}_k + \nabla \mathbf{f}_{k+1} - \nabla \mathbf{f}_k, \qquad (8d)$$

where we have the initial condition $\mathbf{w}_0 = \nabla \mathbf{f}_0$, $\mathbf{y}_0 = \mathbf{1}_n$.

B. Interpretation of ADD-OPT

Based on Eq. (8), we now give an intuitive interpretation on the convergence of ADD-OPT to the optimal solution. By combining Eqs. (8a) and (8d), we obtain that

$$\mathbf{x}_{k+1} = A\mathbf{x}_k - \alpha \left[A\mathbf{w}_{k-1} + \nabla \mathbf{f}_k - \nabla \mathbf{f}_{k-1} \right],$$
$$= A\mathbf{x}_k - \alpha A \left[\frac{A\mathbf{x}_{k-1} - \mathbf{x}_k}{\alpha} \right] - \alpha \left[\nabla \mathbf{f}_k - \nabla \mathbf{f}_{k-1} \right],$$
$$= 2A\mathbf{x}_k - A^2\mathbf{x}_{k-1} - \alpha \left[\nabla \mathbf{f}_k - \nabla \mathbf{f}_{k-1} \right]. \qquad (9)$$

Assume that the sequences generated by Eq. (8) converge to their limits (note that this is not necessarily true), denoted by $\mathbf{x}_\infty$, $\mathbf{y}_\infty$, $\mathbf{w}_\infty$, $\mathbf{z}_\infty$, $\nabla \mathbf{f}_\infty$, respectively. It follows from Eq. (9) that

$$\mathbf{x}_\infty = 2A\mathbf{x}_\infty - A^2\mathbf{x}_\infty - \alpha \left[\nabla \mathbf{f}_\infty - \nabla \mathbf{f}_\infty \right], \qquad (10)$$

which implies that $(I_{np} - A)^2 \mathbf{x}_\infty = \mathbf{0}_{np}$ or $[(I_n - \underline{A})^2 \otimes I_p]\mathbf{x}_\infty = \mathbf{0}_{np}$. Considering that $\mathbf{y}_\infty = \underline{A}\mathbf{y}_\infty$, we obtain that $\mathbf{x}_\infty \in \operatorname{span}\{\mathbf{y}_\infty \otimes \mathbf{u}_p\}$ for some arbitrary p-dimensional vector, $\mathbf{u}_p$. Therefore, it follows that

$$\mathbf{z}_\infty = Y_\infty^{-1}\mathbf{x}_\infty \in \operatorname{span}\{\mathbf{1}_n \otimes \mathbf{u}_p\}, \qquad (11)$$

where $\mathbf{u}_p$ is some arbitrary p-dimensional vector. The consensus is reached.

By summing up the updates in Eq. (9) over k from 0 to ∞, we obtain that

$$\mathbf{x}_\infty = A\mathbf{x}_\infty + \sum_{r=1}^\infty (A - I_{np})\mathbf{x}_r - \sum_{r=0}^\infty (A^2 - A)\mathbf{x}_r - \alpha \nabla \mathbf{f}_\infty.$$

Noting that $\mathbf{x}_\infty = A\mathbf{x}_\infty$, it follows

$$\alpha \nabla \mathbf{f}_\infty = \sum_{r=1}^\infty (A - I_{np})\mathbf{x}_r - \sum_{r=0}^\infty (A^2 - A)\mathbf{x}_r.$$

Therefore, we obtain that

$$\alpha(\mathbf{1}_n \otimes I_p)^\top \nabla \mathbf{f}_\infty$$
$$= \left(\mathbf{1}_n^\top (\underline{A} - I_n) \otimes I_p \right) \sum_{r=1}^\infty \mathbf{x}_r - \left(\mathbf{1}_n^\top (\underline{A}^2 - \underline{A}) \otimes I_p \right) \sum_{r=0}^\infty \mathbf{x}_r,$$
$$= \mathbf{0}_p,$$

which is the optimality condition of Problem P1 considering that $\mathbf{z}_\infty \in \operatorname{span}\{\mathbf{1}_n \otimes \mathbf{u}_p\}$. To summarize, if we assume that the sequences updated in Eq. (8) have limits, $\mathbf{x}_\infty$, $\mathbf{y}_\infty$, $\mathbf{w}_\infty$, $\mathbf{z}_\infty$, $\nabla \mathbf{f}_\infty$, we arrive at a conclusion that $\mathbf{z}_\infty$ achieves consensus and reaches the optimal solution of Problem P1. We next discuss the relations between ADD-OPT and DEXTRA.

C. ADD-OPT and DEXTRA

Recent papers provide a fast distributed algorithm, termed DEXTRA [37, 38], to solve Problem P1 over directed graphs. It achieves a linear convergence rate given that the objective functions are strongly-convex. At the kth iteration of DEXTRA, each agent i keeps and updates three states, $x_{k,i}$, $y_{k,i}$, and $z_{k,i}$. The iteration, in matrix form, is shown as follows.

$$\mathbf{x}_{k+1} = (I_{np} + A)\mathbf{x}_k - \widetilde{A}\mathbf{x}_{k-1} - \alpha \left[\nabla \mathbf{f}_k - \nabla \mathbf{f}_{k-1} \right], \quad (12a)$$
$$\mathbf{y}_{k+1} = \underline{A}\mathbf{y}_k, \qquad (12b)$$
$$\mathbf{z}_{k+1} = Y_{k+1}^{-1}\mathbf{x}_{k+1}, \qquad (12c)$$

where $\widetilde{A}$ is a column-stochastic matrix satisfying that $\widetilde{A} = \theta I_{np} + (1 - \theta)A$ with any $\theta \in (0, \frac{1}{2}]$, and all other notation is the same as from earlier in this paper.

By comparing Eqs. (9) and (12a), (8b) and (12b), and (8c) and (12c), it follows that the only difference between ADD-OPT and DEXTRA lies in the weighting matrices used when updating $\mathbf{x}_k$. From DEXTRA to ADD-OPT, we change $(I_{np} + A)$ in (12a) to $2A$ in (9), and $\widetilde{A}$ to A^2, respectively. Mathematically, if $A = I_{np}$, (equivalently $\underline{A} = I_n$), the two algorithms are the same. With this modification, we will show in Section III that ADD-OPT supports a wider range of step-sizes as compared to DEXTRA, i.e., the greatest lower

bound, $\underline{\alpha}$, of ADD-OPT's step-size is zero while that of DEX-TRA's is strictly positive. This also reveals the reason why in DEXTRA constructing $\underline{A}$ to be an extremely diagonally-dominant matrix is preferred, see Assumption A2(c) in [37]. The more similar $\underline{A}$ is to I_n, the closer $\underline{\alpha}$ approaches zero. However, in DEXTRA, $\underline{\alpha}$ can never reach zero since $\underline{A}$ cannot be the identity, I_n, which otherwise means there is no communication between agents. In Section V, we provide a totally different proof, that is further more compact and elegant when compared to DEXTRA's analysis, to show the linear convergence rate of ADD-OPT.

III. MAIN RESULT

In this section, we analyze ADD-OPT with the help of the following notation. From Eqs. (4)-(7), we further define $\overline{\mathbf{x}}_k$, $\overline{\mathbf{w}}_k$, $\mathbf{z}^*$, $\mathbf{g}_k$, $\mathbf{h}_k \in \mathbb{R}^{np}$ as

$$\overline{\mathbf{x}}_k = \frac{1}{n}(\mathbf{1}_n \otimes I_p)(\mathbf{1}_n^\top \otimes I_p)\mathbf{x}_k, \tag{13}$$

$$\overline{\mathbf{w}}_k = \frac{1}{n}(\mathbf{1}_n \otimes I_p)(\mathbf{1}_n^\top \otimes I_p)\mathbf{w}_k, \tag{14}$$

$$\mathbf{z}^* = \mathbf{1}_n \otimes \underline{\mathbf{z}}^*, \tag{15}$$

$$\mathbf{g}_k = \frac{1}{n}(\mathbf{1}_n \otimes I_p)(\mathbf{1}_n^\top \otimes I_p)\nabla\mathbf{f}_k, \tag{16}$$

$$\mathbf{h}_k = \frac{1}{n}(\mathbf{1}_n \otimes I_p)(\mathbf{1}_n^\top \otimes I_p)\nabla\mathbf{f}(\overline{\mathbf{x}}_k), \tag{17}$$

where

$$\nabla\mathbf{f}(\overline{\mathbf{x}}_k) = \begin{bmatrix} \nabla f_1(\frac{1}{n}(\mathbf{1}_n^\top \otimes I_p)\mathbf{x}_k) \\ \vdots \\ \nabla f_n(\frac{1}{n}(\mathbf{1}_n^\top \otimes I_p)\mathbf{x}_k) \end{bmatrix},$$

stacks its components in a column. We denote constants, τ, ϵ, and η as

$$\tau = \|A - I_{np}\|_2, \tag{18}$$

$$\epsilon = \|I_{np} - A_\infty\|_2, \tag{19}$$

$$\eta = \max\left(|1 - n\alpha l|, |1 - n\alpha s|\right), \tag{20}$$

where A is the column-stochastic weight matrix used in Eq. (8), $A_\infty = \lim_{k\to\infty} A^k$ represents A's limit, α is the step-size, and l and s are respectively Lipschitz and strong-convexity constants from Assumption A2. Let Y_∞ be the limit of Y_k in Eq. (7),

$$Y_\infty = \lim_{k\to\infty} Y_k, \tag{21}$$

and y and y_- be the supremum of $\|Y_k\|_2$ and $\|Y_k^{-1}\|_2$ over k, respectively, i.e.,

$$y = \sup_k \|Y_k\|_2, \tag{22}$$

$$y_- = \sup_k \|Y_k^{-1}\|_2. \tag{23}$$

Note that the existence of the limits, A_∞ and Y_∞, will be clear in the following lemmas. Moreover, we define two constants, σ, and, γ_1, through the following two lemmas, which are related to the convergence of A and Y_∞.

Lemma 1. *(Nedic et al. [26]) Let Assumption A1 hold. Consider Y_k and its limit Y_∞ as defined before. There exist $0 < \gamma_1 < 1$ and $0 < T < \infty$ such that for all k*

$$\|Y_k - Y_\infty\|_2 \leq T\gamma_1^k. \tag{24}$$

Lemma 2. *Let Assumption A1 hold. Consider Y_∞ in Eq. (21) with A being the column-stochastic matrix used in Eq. (8). For any $\mathbf{a} \in \mathbb{R}^{np}$, define $\overline{\mathbf{a}} = \frac{1}{n}(\mathbf{1}_n \otimes I_p)(\mathbf{1}_n^\top \otimes I_p)\mathbf{a}$. Then, there exists $0 < \sigma < 1$ such that for all k*

$$\|A\mathbf{a} - Y_\infty\overline{\mathbf{a}}\| \leq \sigma\|\mathbf{a} - Y_\infty\overline{\mathbf{a}}\|. \tag{25}$$

Proof. First note that $A = \underline{A} \otimes I_p$. Since $\underline{A}$ is irreducible, column-stochastic with positive diagonals, from Perron-Frobenius theorem we note that $\rho(\underline{A}) = 1$, every eigenvalue of $\underline{A}$ other than 1 is strictly less than $\rho(\underline{A})$, and $\boldsymbol{\pi}$ is a strictly positive (right) eigenvector corresponding to the eigenvalue of 1 such that $\mathbf{1}_n^\top\boldsymbol{\pi} = 1$; thus $\lim_{k\to\infty} \underline{A}^k = \boldsymbol{\pi}\mathbf{1}_n^\top$. Recalling Eq. (6), we have $A_\infty = \lim_{k\to\infty} A^k = \lim_{k\to\infty}(\underline{A} \otimes I_p)^k = (\lim_{k\to\infty} \underline{A}^k) \otimes I_p = (\boldsymbol{\pi}\mathbf{1}_n^\top) \otimes I_p$. It follows that:

$$AA_\infty = (\underline{A} \otimes I_p)\left((\boldsymbol{\pi}\mathbf{1}_n^\top) \otimes I_p\right) = (\underline{A}\boldsymbol{\pi}\mathbf{1}_n^\top) \otimes I_p = A_\infty;$$

$$A_\infty A_\infty = \left((\boldsymbol{\pi}\mathbf{1}_n^\top) \otimes I_p\right)\left((\boldsymbol{\pi}\mathbf{1}_n^\top) \otimes I_p\right),$$
$$= (\boldsymbol{\pi}\mathbf{1}_n^\top\boldsymbol{\pi}\mathbf{1}_n^\top) \otimes I_p = A_\infty.$$

Thus $AA_\infty - A_\infty A_\infty$ is a zero matrix. It can also be verified that $\frac{1}{n}Y_\infty(\mathbf{1}_n \otimes I_p)(\mathbf{1}_n^\top \otimes I_p) = A_\infty$. Based on the discussion above, we have

$$A\mathbf{a} - Y_\infty\overline{\mathbf{a}} = (A - A_\infty)(\mathbf{a} - A_\infty\mathbf{a}) = (A - A_\infty)(\mathbf{a} - Y_\infty\overline{\mathbf{a}}).$$

Next we note that

$$\rho(A - A_\infty) = \rho\left((\underline{A} - \boldsymbol{\pi}\mathbf{1}_n^\top) \otimes I_p\right) = \rho(\underline{A} - \boldsymbol{\pi}\mathbf{1}_n^\top) < 1,$$

and there exists a matrix norm such that $\|A - A_\infty\| < 1$ with a compatible vector norm, $\|\cdot\|$, see [41]: Chapter 5 for details, i.e.,

$$\|A\mathbf{a} - Y_\infty\overline{\mathbf{a}}\| \leq \|A - A_\infty\|\|\mathbf{a} - Y_\infty\overline{\mathbf{a}}\|, \tag{26}$$

and the lemma follows with $\sigma = \|A - A_\infty\|$. $\square$

Based on the above notation, we finally denote $\mathbf{t}_k$, $\mathbf{s}_k \in \mathbb{R}^3$, and G, $H_k \in \mathbb{R}^{3\times3}$, for all k as

$$\mathbf{t}_k = \begin{bmatrix} \|\mathbf{x}_k - Y_\infty\overline{\mathbf{x}}_k\| \\ \|\overline{\mathbf{x}}_k - \mathbf{z}^*\|_2 \\ \|\mathbf{w}_k - Y_\infty\mathbf{g}_k\| \end{bmatrix}, \qquad \mathbf{s}_k = \begin{bmatrix} \|\mathbf{x}_k\|_2 \\ 0 \\ 0 \end{bmatrix},$$

$$G = \begin{bmatrix} \sigma & 0 & \alpha \\ \alpha cly_- & \eta & 0 \\ cd\epsilon ly_-(\tau + \alpha lyy_-) & \alpha d\epsilon l^2 yy_- & \sigma + \alpha cd\epsilon ly_- \end{bmatrix},$$

$$H_k = \begin{bmatrix} 0 & 0 & 0 \\ \alpha ly_- T\gamma_1^{k-1} & 0 & 0 \\ (\alpha ly + 2)d\epsilon ly_-^2 T\gamma_1^{k-1} & 0 & 0 \end{bmatrix}. \tag{27}$$

We now state a key relation of this paper.

Lemma 3. *Let the directed graph be strongly-connected and the optimal solution of Problem P1 exist (Assumption A1 and A3). Let $\mathbf{t}_k$, $\mathbf{s}_k$, G, and H_k be defined in Eq. (27), in which $\mathbf{x}_k$ is the sequence generated by ADD-OPT, Eq. (8), over k. Under the smooth and strong-convexity assumptions (Assumption A2), we have $\mathbf{t}_k$, $\mathbf{s}_k$, G, and H_k satisfy the following linear relation,*

$$\mathbf{t}_k \leq G\mathbf{t}_{k-1} + H_{k-1}\mathbf{s}_{k-1}. \tag{28}$$

Proof. See Section V. □

We leave the complete proof to Section V, with the help of several auxiliary relations in Section IV. Note that Eq. (28) provides a linear iterative relation between $\mathbf{t}_k$ and $\mathbf{t}_{k-1}$ with matrices, G and H_k. Thus, the convergence of $\mathbf{t}_k$ is fully determined by G and H_k. More specifically, if we want to prove linear convergence of $\|\mathbf{t}_k\|_2$ to zero, it is sufficient to show that $\rho(G) < 1$, where $\rho(\cdot)$ denotes the spectral radius, as well as the linear decaying of H_k, which is straightforward since $0 < \gamma_1 < 1$. In Lemma 4, we first show that with appropriate step-size, the spectral radius of G is less than 1. Afterwards, in Lemma 5, we study the convergence properties of the matrices involving G and H_k.

Lemma 4. *Consider the matrix G defined in Eq. (27) as a function of the step-size, α, denoted in this lemma as G_α to motivate this dependence. It follows that $\rho(G_\alpha) < 1$ if the step-size, $\alpha \in (0, \overline{\alpha})$, where*

$$\overline{\alpha} = \min\left\{\frac{\sqrt{\Delta^2 + 4ns(1-\sigma)^2 cd\epsilon l^2 yy_-^2(l+ns)} - \Delta}{2cd\epsilon l^2 yy_-^2(l+ns)}, \frac{1}{nl}\right\}, \tag{29}$$

and $\Delta = nscd\epsilon ly_-(1-\sigma+\tau)$, where c and d are the constants from the equivalence of $\|\cdot\|$ defined in Lemma 2 and $\|\cdot\|_2$.

Proof. First, if $\alpha < \frac{1}{nl}$ then $\eta = 1 - \alpha ns$, since $l \geq s$ (see e.g., [42]: Chapter 3 for details). When $\alpha = 0$, we have that

$$G_0 = \begin{bmatrix} \sigma & 0 & 0 \\ 0 & 1 & 0 \\ cd\epsilon l\tau y_- & 0 & \sigma \end{bmatrix}, \tag{30}$$

the eigenvalues of which are σ, σ, and 1. Hence, $\rho(G_0) = 1$. We now consider how the eigenvalue of 1 is changed if we slightly increase α from 0. Let $\mathcal{P}_{G_\alpha}(q) = \det(qI_n - G_\alpha)$, i.e., the characteristic polynomial of G_α. Setting $\det(qI_n - G_\alpha) = 0$, we get the following equation.

$$((q-\sigma)^2 - \alpha cd\epsilon ly_-(q-\sigma))(q-1+n\alpha s) - \alpha^3 cd\epsilon l^3 yy_-^2$$
$$-\alpha(q-1+n\alpha s)(cd\epsilon l\tau y_- + \alpha(cd\epsilon l^2 yy_-^2)) = 0. \tag{31}$$

Since we have already shown that 1 is one of the eigenvalues of G_0, Eq. (31) holds when $q = 1$ and $\alpha = 0$. By taking the derivative on both sides of Eq. (31), with $q = 1$ and $\alpha = 0$, we obtain that $\frac{dq}{d\alpha}|_{\alpha=0,q=1} = -ns < 0$. This leads to the fact that when α slightly increases from 0, $\rho(G_\alpha) < 1$ since the

eigenvalues are continuous functions of the parameters of a matrix.

We next calculate all possible values of α for which G_α has an eigenvalue of 1. Let $q = 1$ in Eq. (31) and solve for the step-size, α; we obtain three solutions: $\alpha_1 = 0$, $\alpha_2 < 0$, and

$$\alpha_3 = \frac{\sqrt{\Delta^2 + 4ns(1-\sigma)^2 cd\epsilon l^2 yy_-^2(l+ns)} - \Delta}{2cd\epsilon l^2 yy_-^2(l+ns)} > 0.$$

Since there are no other values of α with which G_α has an eigenvalue of 1, all eigenvalues of G_α are less than 1, i.e., $\rho(G_\alpha) < 1$, when $\alpha \in (0, \overline{\alpha})$. □

We note that $\overline{\alpha}$ depends on the global knowledge and it may not be possible to precisely compute it in a distributed fashion. However, this value may be estimated as we will show in Section VI, see e.g., [16], for a similar approach.

Lemma 5. *With the step-size, $\alpha \in (0, \overline{\alpha})$, where $\overline{\alpha}$ is defined in Eq. (29), the following statements hold: $\forall k$,*

(a) there exists $0 < \gamma_1 < 1$ and $0 < \Gamma_1 < \infty$, where γ_1 is defined in Eq. (24), such that

$$\|H_k\|_2 = \Gamma_1\gamma_1^k;$$

(b) there exists $0 < \gamma_2 < 1$ and $0 < \Gamma_2 < \infty$, such that

$$\|G^k\|_2 \leq \Gamma_2\gamma_2^k;$$

(c) let $\gamma = \max\{\gamma_1, \gamma_2\}$ and $\Gamma = \Gamma_1\Gamma_2/\gamma$, such that for all $0 \leq r \leq k-1$,

$$\|G^{k-r-1}H_r\|_2 \leq \Gamma\gamma^k.$$

Proof.

(a) This can be verified according to Eq. (27) and by letting

$$\Gamma_1 = \frac{1}{\gamma_1}\sqrt{(\alpha ly_-T)^2 + (\alpha yl + 2)^2(d\epsilon ly_-^2 T)^2}.$$

(b) Note that $\rho(G) < 1$ when $\alpha \in (0, \overline{\alpha})$. Therefore, the value of some matrix norm of G, denoted by γ_2, is strictly less than 1. Since all matrix norms are equivalent, we have $\|G^k\|_2 \leq \Gamma_2\gamma_2^k$, for some positive constant Γ_2.

(c) The proof of (c) is achieved by combining (a) and (b). □

Lemma 6. *(Polyak [43]) If nonnegative sequences $\{v_k\}$, $\{u_k\}$, $\{b_k\}$ and $\{c_k\}$ are such that $\sum_{k=0}^{\infty} b_k < \infty$, $\sum_{k=0}^{\infty} c_k < \infty$ and*

$$v_{k+1} \leq (1+b_k)v_k - u_k + c_k, \quad \forall t \geq 0,$$

then $\{v_k\}$ converges and $\sum_{k=0}^{\infty} u_k < \infty$.

We now present the main result of this paper in Theorem 1, which shows the linear convergence rate of ADD-OPT.

Theorem 1. *Let the Assumptions A1-A3 hold. With the step-size, $\alpha \in (0, \overline{\alpha})$, where $\overline{\alpha}$ is defined in Eq. (29), the sequence, $\{\mathbf{z}_k\}$, generated by ADD-OPT, converges exactly to*

the unique optimizer, $\mathbf{z}^*$, at a linear rate, i.e., there exist some positive constant $M > 0$, such that for any k,

$$\|\mathbf{z}_k - \mathbf{z}^*\|_2 \leq M(\gamma + \xi)^k, \tag{32}$$

where γ is used in Lemma 5(c) and ξ is a arbitrarily small constant.

Proof. We write Eq. (28) recursively, leading to

$$\mathbf{t}_k \leq G^k \mathbf{t}_0 + \sum_{r=0}^{k-1} G^{k-r-1} H_r \mathbf{s}_r. \tag{33}$$

By taking the norm on both sides of Eq. (33) and considering Lemma 5, we obtain that

$$\|\mathbf{t}_k\|_2 \leq \|G^k\|_2 \|\mathbf{t}_0\|_2 + \sum_{r=0}^{k-1} \|G^{k-r-1} H_r\|_2 \|\mathbf{s}_r\|_2,$$

$$\leq \Gamma_2 \gamma_2^k \|\mathbf{t}_0\|_2 + \sum_{r=0}^{k-1} \Gamma \gamma^k \|\mathbf{s}_r\|_2, \tag{34}$$

in which we can bound $\|\mathbf{s}_r\|_2$ as

$$\|\mathbf{s}_r\|_2 \leq \|\mathbf{x}_r - Y_\infty \overline{\mathbf{x}}_r\|_2 + \|Y_\infty\|_2 \|\overline{\mathbf{x}}_r - \mathbf{z}^*\|_2 + \|Y_\infty\|_2 \|\mathbf{z}^*\|_2,$$
$$\leq (c + y) \|\mathbf{t}_r\|_2 + y \|\mathbf{z}^*\|_2. \tag{35}$$

Therefore, we have that for all k

$$\|\mathbf{t}_k\|_2 \leq \left(\Gamma_2 \|\mathbf{t}_0\|_2 + \Gamma(c + y) \sum_{r=0}^{k-1} \|\mathbf{t}_r\|_2 + \Gamma y k \|\mathbf{z}^*\|_2 \right) \gamma^k. \tag{36}$$

Denote $v_k = \sum_{r=0}^{k-1} \|\mathbf{t}_r\|_2$, $s_k = \Gamma_2 \|\mathbf{t}_0\|_2 + \Gamma y k \|\mathbf{z}^*\|_2$, and $b = \Gamma(c + y)$, then Eq. (36) can be written as

$$\|\mathbf{t}_k\|_2 = v_{k+1} - v_k \leq (s_k + b v_k) \gamma^k, \tag{37}$$

which implies that $v_{k+1} \leq (1 + b\gamma^k) v_k + s_k \gamma^k$. Applying Lemma 6 with $b_k = b\gamma^k$ and $c_k = s_k \gamma^k$ (here $u_k = 0$), we have that v_k converges[2]. and therefore is bounded. By Eq. (37), $\forall \mu \in (\gamma, 1)$ we have

$$\lim_{k \to \infty} \frac{\|\mathbf{t}_k\|_2}{\mu^k} \leq \lim_{k \to \infty} \frac{(s_k + b v_k) \gamma^k}{\mu^k} = 0. \tag{38}$$

Therefore, $\|\mathbf{t}_k\|_2 = O(\mu^k)$. In other words, there exists some positive constant Φ such that for all k, we have:

$$\|\mathbf{t}_k\|_2 \leq \Phi(\gamma + \xi)^k, \tag{39}$$

where ξ is a arbitrarily small constant. Moreover, $\|\mathbf{z}_k - \mathbf{z}^*\|_2$ and $\|\mathbf{t}_k\|_2$ satisfy the relation that

$$\|\mathbf{z}_k - \mathbf{z}^*\|_2 \leq \left\| Y_k^{-1} \mathbf{x}_k - Y_k^{-1} Y_\infty \overline{\mathbf{x}}_k \right\|_2 + \left\| Y_k^{-1} Y_\infty \mathbf{z}^* - \mathbf{z}^* \right\|_2$$
$$+ \left\| Y_k^{-1} Y_\infty \overline{\mathbf{x}}_k - Y_k^{-1} Y_\infty \mathbf{z}^* \right\|_2,$$
$$\leq y_-(c + y) \|\mathbf{t}_k\|_2 + y_- T \gamma_1^k \|\mathbf{z}^*\|_2, \tag{40}$$

where in the second inequality we use the relation

$$\|Y_k^{-1} Y_\infty - I_{np}\|_2 \leq \|Y_k^{-1}\|_2 \|Y_\infty - Y_k\|_2 \leq y_- T \gamma_1^k,$$

[2]In order to apply Lemma 6, we need to show that $\sum_{k=0}^\infty s_k \gamma^k < \infty$, which follows from the fact that $\lim_{k \to \infty} \frac{s_{k+1} \gamma^{k+1}}{s_k \gamma^k} = \gamma < 1$.

achieved from Eq. (24). By combining Eqs. (39) and (40), we obtain that

$$\|\mathbf{z}_k - \mathbf{z}^*\|_2 \leq \left(y_-(c + y)\Phi + y_- T \|\mathbf{z}^*\|_2 \right)(\gamma + \xi)^k,$$

where ξ is a arbitrarily small constant. The proof of theorem is completed by letting $M = y_-(c + y)\Phi + y_- T \|\mathbf{z}^*\|_2$. $\square$

Theorem 1 shows the linear convergence rate of ADD-OPT. Although ADD-OPT works for a small enough step-size, how small is sufficient may require some estimation of the upper bound, which we discuss this in Section VI. This notion of sufficiently small step-sizes is not uncommon in the literature, see e.g., [10, 26]. Next, each agent must agree on the same value of step-size that may be pre-programmed to avoid implementing an agreement protocol. We now prove Lemma 3 in Sections IV and V.

IV. Auxiliary Relations

We provide several basic relations in this section, which will help the proof of Lemma 3. Lemma 7 derives iterative equations that govern the average sequences, $\overline{\mathbf{x}}_k$ and $\overline{\mathbf{w}}_k$. Lemma 8 gives inequalities that are direct consequences of Eq. (24). Lemma 9 can be found in the standard optimization literature, see e.g., [42]. It states that if we perform a gradient-descent step with a fixed step-size for a smooth, strongly-convex function, then the distance to optimizer shrinks by at least a fixed ratio.

Lemma 7. *Recall $\overline{\mathbf{x}}_k$ from Eq. (13) and $\overline{\mathbf{w}}_k$ from Eq. (14). The following equations hold for all k,*

(a) $\overline{\mathbf{w}}_k = \mathbf{g}_k$;
(b) $\overline{\mathbf{x}}_{k+1} = \overline{\mathbf{x}}_k - \alpha \mathbf{g}_k$.

Proof. Since A is column-stochastic, satisfying $(\mathbf{1}_n^\top \otimes I_p) A = \mathbf{1}_n^\top \otimes I_p$, we obtain that

$$\overline{\mathbf{w}}_k = \frac{1}{n}(\mathbf{1}_n \otimes I_p)(\mathbf{1}_n^\top \otimes I_p)(A\mathbf{w}_{k-1} + \nabla \mathbf{f}_k - \nabla \mathbf{f}_{k-1}),$$
$$= \overline{\mathbf{w}}_{k-1} + \mathbf{g}_k - \mathbf{g}_{k-1}.$$

Do this recursively, and we have that

$$\overline{\mathbf{w}}_k = \overline{\mathbf{w}}_0 + \mathbf{g}_k - \mathbf{g}_0.$$

Recall that we have the initial condition that $\mathbf{w}_0 = \nabla \mathbf{f}_0$, which is equivalent to $\overline{\mathbf{w}}_0 = \mathbf{g}_0$. Hence, we achieve the result of (a). The proof of (b) is obtained by the following derivation,

$$\overline{\mathbf{x}}_{k+1} = \frac{1}{n}(\mathbf{1}_n \otimes I_p)(\mathbf{1}_n^\top \otimes I_p)(A\mathbf{x}_k - \alpha \mathbf{w}_k)$$
$$= \overline{\mathbf{x}}_k - \alpha \overline{\mathbf{w}}_k,$$
$$= \overline{\mathbf{x}}_k - \alpha \mathbf{g}_k,$$

where the last equation uses the result of (a). $\square$

Lemma 8. *Recall Lemma 1, Y_k from Eq. (7), and Y_∞ from Eq. (21). The following inequalities hold for all $k \geq 1$,*

(a) $\left\| Y_{k-1}^{-1} Y_\infty - I_{np} \right\|_2 \leq y_- T \gamma_1^{k-1}$;
(b) $\left\| Y_k^{-1} - Y_{k-1}^{-1} \right\|_2 \leq 2y_-^2 T \gamma_1^{k-1},$
where y_- is defined in Eq. (23).

Proof. By considering Eq. (24), it follows that

$$\left\|Y_{k-1}^{-1}Y_\infty - I_{np}\right\|_2 \leq \left\|Y_{k-1}^{-1}\right\|_2 \left\|Y_\infty - Y_{k-1}\right\|_2 \leq y_- T\gamma_1^{k-1}.$$

The proof of (b) follows by

$$\begin{aligned}
\left\|Y_k^{-1} - Y_{k-1}^{-1}\right\|_2 &\leq \left\|Y_{k-1}^{-1}\right\|_2 \left\|Y_{k-1} - Y_k\right\|_2 \left\|Y_k^{-1}\right\|_2, \\
&\leq 2y_-^2 T\gamma_1^{k-1},
\end{aligned}$$

which completes the proof. $\qquad\square$

Lemma 9. *(Bubeck [42]) Let Assumption A2 hold for the objective functions, $f_i(\mathbf{z})$, in Problem P1, and let s and l be the strong-convexity and Lipschitz-continuity constants, respectively. For any $\mathbf{z} \in \mathbb{R}^p$, define $\mathbf{z}_+ = \mathbf{z} - \alpha\nabla\mathbf{f}(\mathbf{z})$, where $0 < \alpha < \frac{2}{nl}$. Then*

$$\left\|\mathbf{z}_+ - \underline{\mathbf{z}}^*\right\|_2 \leq \eta\left\|\mathbf{z} - \underline{\mathbf{z}}^*\right\|_2,$$

where $\eta = \max\left(|1 - \alpha nl|, |1 - \alpha ns|\right)$.

V. CONVERGENCE ANALYSIS

We now provide the proof of Lemma 3. We will bound $\|\mathbf{x}_k - Y_\infty\overline{\mathbf{x}}_k\|$, $\|\overline{\mathbf{x}}_k - \mathbf{z}^*\|_2$, and $\|\mathbf{w}_k - Y_\infty\mathbf{g}_k\|$, linearly in terms of their past values, i.e., $\|\mathbf{x}_{k-1} - Y_\infty\overline{\mathbf{x}}_{k-1}\|$, $\|\overline{\mathbf{x}}_{k-1} - \mathbf{z}^*\|_2$, and $\|\mathbf{w}_{k-1} - Y_\infty\mathbf{g}_{k-1}\|$, as well as $\|\mathbf{x}_{k-1}\|_2$. The coefficients are the entries of G and H_{k-1}.

Step 1: Bound $\|\mathbf{x}_k - Y_\infty\overline{\mathbf{x}}_k\|$.
According to Eq. (8a) and Lemma 7(b), we obtain that

$$\begin{aligned}
\|\mathbf{x}_k - Y_\infty\overline{\mathbf{x}}_k\| &\leq \|A\mathbf{x}_{k-1} - Y_\infty\overline{\mathbf{x}}_{k-1}\| \\
&\quad + \alpha\|\mathbf{w}_{k-1} - Y_\infty\mathbf{g}_{k-1}\|.
\end{aligned} \tag{41}$$

Noticing that $\|A\mathbf{x}_{k-1} - Y_\infty\overline{\mathbf{x}}_{k-1}\| \leq \sigma\|\mathbf{x}_{k-1} - Y_\infty\overline{\mathbf{x}}_{k-1}\|$ from Eq. (25), we have

$$\begin{aligned}
\|\mathbf{x}_k - Y_\infty\overline{\mathbf{x}}_k\| &\leq \sigma\|\mathbf{x}_{k-1} - Y_\infty\overline{\mathbf{x}}_{k-1}\| \\
&\quad + \alpha\|\mathbf{w}_{k-1} - Y_\infty\mathbf{g}_{k-1}\|.
\end{aligned} \tag{42}$$

Step 2: Bound $\|\overline{\mathbf{x}}_k - \mathbf{z}^*\|_2$.
By considering Lemma 7(b), we obtain that

$$\overline{\mathbf{x}}_k = [\overline{\mathbf{x}}_{k-1} - \alpha\mathbf{h}_{k-1}] - \alpha[\mathbf{g}_{k-1} - \mathbf{h}_{k-1}]. \tag{43}$$

Let $\mathbf{x}_+ = \overline{\mathbf{x}}_{k-1} - \alpha\mathbf{h}_{k-1}$, which is a (centralized) gradient-descent step with respect to the global objective function in Problem P1. Therefore, from Lemma 9,

$$\|\mathbf{x}_+ - \mathbf{z}^*\|_2 \leq \eta\|\overline{\mathbf{x}}_{k-1} - \mathbf{z}^*\|_2. \tag{44}$$

From the Lipschitz-continuity, Assumption A2(a), we obtain

$$\|\mathbf{g}_{k-1} - \mathbf{h}_{k-1}\|_2 \leq \left\|\frac{1}{n}(\mathbf{1}_n\mathbf{1}_n^\top)\otimes I_p\right\|_2 l\|\mathbf{z}_{k-1} - \overline{\mathbf{x}}_{k-1}\|_2. \tag{45}$$

Therefore, it follows that

$$\begin{aligned}
\|\overline{\mathbf{x}}_k - \mathbf{z}^*\|_2 &\leq \|\mathbf{x}_+ - \mathbf{z}^*\|_2 + \alpha\|\mathbf{g}_{k-1} - \mathbf{h}_{k-1}\|_2, \\
&\leq \eta\|\overline{\mathbf{x}}_{k-1} - \mathbf{z}^*\|_2 + \alpha l\|\mathbf{z}_{k-1} - \overline{\mathbf{x}}_{k-1}\|_2.
\end{aligned} \tag{46}$$

From Eq. (8c) and Lemma 8(a), it follows that

$$\begin{aligned}
\|\mathbf{z}_{k-1} - \overline{\mathbf{x}}_{k-1}\|_2 &\leq \left\|Y_{k-1}^{-1}\left(\mathbf{x}_{k-1} - Y_\infty\overline{\mathbf{x}}_{k-1}\right)\right\|_2 \\
&\quad + \left\|\left(Y_{k-1}^{-1}Y_\infty - I_{np}\right)\overline{\mathbf{x}}_{k-1}\right\|_2,
\end{aligned}$$

$$\begin{aligned}
&\leq y_-\|\mathbf{x}_{k-1} - Y_\infty\overline{\mathbf{x}}_{k-1}\|_2 \\
&\quad + y_- T\gamma_1^{k-1}\|\mathbf{x}_{k-1}\|_2,
\end{aligned} \tag{47}$$

where in the second inequality we also make use of the relation $\|\overline{\mathbf{x}}_{k-1}\|_2 \leq \|\mathbf{x}_{k-1}\|_2$. By substituting Eq. (47) into Eq. (46), we obtain that

$$\begin{aligned}
\|\overline{\mathbf{x}}_k - \mathbf{z}^*\|_2 &\leq \alpha cl y_-\|\mathbf{x}_{k-1} - Y_\infty\overline{\mathbf{x}}_{k-1}\| + \eta\|\overline{\mathbf{x}}_{k-1} - \mathbf{z}^*\|_2 \\
&\quad + \alpha l y_- T\gamma_1^{k-1}\|\mathbf{x}_{k-1}\|_2.
\end{aligned} \tag{48}$$

Step 3: Bound $\|\mathbf{w}_k - Y_\infty\mathbf{g}_k\|$.
According to Eq. (8d), we have

$$\begin{aligned}
\|\mathbf{w}_k - Y_\infty\mathbf{g}_k\| &\leq \|A\mathbf{w}_{k-1} - Y_\infty\mathbf{g}_{k-1}\| \\
&\quad + \|(\nabla\mathbf{f}_k - \nabla\mathbf{f}_{k-1}) - (Y_\infty\mathbf{g}_k - Y_\infty\mathbf{g}_{k-1})\|.
\end{aligned}$$

With Lemma 7(a) and Eq. (25), we obtain that

$$\begin{aligned}
\|A\mathbf{w}_{k-1} - Y_\infty\mathbf{g}_{k-1}\| &= \|A\mathbf{w}_{k-1} - Y_\infty\overline{\mathbf{w}}_{k-1}\|, \\
&\leq \sigma\|\mathbf{w}_{k-1} - Y_\infty\overline{\mathbf{w}}_{k-1}\|.
\end{aligned} \tag{49}$$

It follows from the definition of $\mathbf{g}_k$ that

$$\begin{aligned}
&\|(\nabla\mathbf{f}_k - \nabla\mathbf{f}_{k-1}) - (Y_\infty\mathbf{g}_k - Y_\infty\mathbf{g}_{k-1})\|_2 \\
&= \left\|\left(I_{np} - \frac{1}{n}Y_\infty(\mathbf{1}_n\otimes I_p)(\mathbf{1}_n^\top\otimes I_p)\right)(\nabla\mathbf{f}_k - \nabla\mathbf{f}_{k-1})\right\|_2.
\end{aligned} \tag{50}$$

Since $\frac{1}{n}Y_\infty(\mathbf{1}_n\otimes I_p)(\mathbf{1}_n^\top\otimes I_p) = A_\infty$, we obtain that

$$\|(\nabla\mathbf{f}_k - \nabla\mathbf{f}_{k-1}) - (Y_\infty\mathbf{g}_k - Y_\infty\mathbf{g}_{k-1})\|_2 \leq \epsilon l\|\mathbf{z}_k - \mathbf{z}_{k-1}\|_2,$$

where we use the Lipschitz-continuity, Assumption A2(a). Therefore, we have

$$\begin{aligned}
\|\mathbf{w}_k - Y_\infty\mathbf{g}_k\| &\leq \sigma\|\mathbf{w}_{k-1} - Y_\infty\mathbf{g}_{k-1}\| \\
&\quad + d\epsilon l\|\mathbf{z}_k - \mathbf{z}_{k-1}\|_2.
\end{aligned} \tag{51}$$

We now bound $\|\mathbf{z}_k - \mathbf{z}_{k-1}\|_2$. Note that

$$\begin{aligned}
\|\mathbf{h}_{k-1}\|_2 &= \left\|\frac{1}{n}(\mathbf{1}_n\otimes I_p)(\mathbf{1}_n^\top\otimes I_p)\nabla\mathbf{f}(\overline{\mathbf{x}}_{k-1})\right\|_2 \\
&\leq l\|\overline{\mathbf{x}}_{k-1} - \mathbf{z}^*\|_2.
\end{aligned} \tag{52}$$

As a result, we have

$$\begin{aligned}
\left\|Y_k^{-1}\mathbf{w}_{k-1}\right\|_2 &\leq \left\|Y_k^{-1}\left(\mathbf{w}_{k-1} - Y_\infty\mathbf{g}_{k-1}\right)\right\|_2 \\
&\quad + \left\|Y_k^{-1}Y_\infty\mathbf{h}_{k-1}\right\|_2 \\
&\quad + \left\|Y_k^{-1}Y_\infty\left(\mathbf{g}_{k-1} - \mathbf{h}_{k-1}\right)\right\|_2, \\
&\leq y_-\|\mathbf{w}_{k-1} - Y_\infty\mathbf{g}_{k-1}\|_2 \\
&\quad + y_- yl\|\overline{\mathbf{x}}_{k-1} - \mathbf{z}^*\|_2 \\
&\quad + y_- yl\|\mathbf{z}_{k-1} - \overline{\mathbf{x}}_{k-1}\|_2, \\
&\leq y_-\|\mathbf{w}_{k-1} - Y_\infty\mathbf{g}_{k-1}\|_2 \\
&\quad + y_- yl\|\overline{\mathbf{x}}_{k-1} - \mathbf{z}^*\|_2 \\
&\quad + y_-^2 yl\|\mathbf{x}_{k-1} - Y_\infty\overline{\mathbf{x}}_{k-1}\|_2 \\
&\quad + y_-^2 ylT\gamma_1^{k-1}\|\mathbf{x}_{k-1}\|_2,
\end{aligned} \tag{53}$$

where the last inequality holds due to Eq. (47). With the upper bound of $\|Y_k^{-1}\mathbf{w}_{k-1}\|_2$ provided in the preceding relation and

the equality that $(A - I_{np})Y_\infty \overline{\mathbf{x}}_{k-1} = \mathbf{0}_n$, we can bound $\|\mathbf{z}_k - \mathbf{z}_{k-1}\|_2$ as follows.

$$\begin{aligned}
\|\mathbf{z}_k - \mathbf{z}_{k-1}\|_2 &\leq \left\|Y_k^{-1}\left(\mathbf{x}_k - \mathbf{x}_{k-1}\right)\right\|_2 \\
&\quad + \left\|\left(Y_k^{-1} - Y_{k-1}^{-1}\right)\mathbf{x}_{k-1}\right\|_2, \\
&\leq \left\|Y_k^{-1}\left(A - I_{np}\right)\mathbf{x}_{k-1}\right\|_2 + \alpha\left\|Y_k^{-1}\mathbf{w}_{k-1}\right\|_2 \\
&\quad + \left\|Y_k^{-1} - Y_{k-1}^{-1}\right\|_2 \|\mathbf{x}_{k-1}\|_2, \\
&\leq (y_- \tau + \alpha y_-^2 yl)\|\mathbf{x}_{k-1} - Y_\infty \overline{\mathbf{x}}_{k-1}\|_2 \\
&\quad + \alpha y_- \|\mathbf{w}_{k-1} - Y_\infty \mathbf{g}_{k-1}\|_2 \\
&\quad + \alpha y_- yl \|\overline{\mathbf{x}}_{k-1} - \mathbf{z}^*\|_2 \\
&\quad + (\alpha yl + 2)y_-^2 T\gamma_1^{k-1}\|\mathbf{x}_{k-1}\|_2.
\end{aligned} \tag{54}$$

By substituting Eq. (54) in Eq. (51), we obtain that

$$\begin{aligned}
\|\mathbf{w}_k - Y_\infty \mathbf{g}_k\| &\leq (cd\epsilon l\tau y_- + \alpha cd\epsilon l^2 yy_-^2)\|\mathbf{x}_{k-1} - Y_\infty \overline{\mathbf{x}}_{k-1}\| \\
&\quad + \alpha d\epsilon l^2 yy_- \|\overline{\mathbf{x}}_{k-1} - \mathbf{z}^*\|_2 \\
&\quad + (\sigma + \alpha cd\epsilon ly_-)\|\mathbf{w}_{k-1} - Y_\infty \mathbf{g}_{k-1}\| \\
&\quad + (\alpha yl + 2)d\epsilon ly_-^2 T\gamma_1^{k-1}\|\mathbf{x}_{k-1}\|_2.
\end{aligned} \tag{55}$$

Step 4: By combining Eqs. (42) in step 1, (48) in step 2, and (55) in step 3, we complete the proof.

VI. NUMERICAL EXPERIMENTS

In this section, we analyze the performance of ADD-OPT. Our numerical experiments are based on the distributed logistic regression problem over a directed graph:

$$\mathbf{z}^* = \underset{\mathbf{z} \in \mathbb{R}^p}{\operatorname{argmin}}\left(\frac{\beta}{2}\|\mathbf{z}\|_2^2 + \sum_{i=1}^{n}\sum_{j=1}^{m_i}\ln\left[1 + \exp\left(-b_{ij}\mathbf{c}_{ij}^\top \mathbf{z}\right)\right]\right).$$

Each agent i has access to m_i training examples, $(\mathbf{c}_{ij}, b_{ij}) \in \mathbb{R}^p \times \{-1, +1\}$, where $\mathbf{c}_{ij}$ includes the p features of the jth training example of agent i and b_{ij} is the corresponding label. This problem can be formulated in the form of Problem P1 with the local objective function, f_i, being

$$f_i = \frac{\beta}{2n}\|\mathbf{z}\|_2^2 + \sum_{j=1}^{m_i}\ln\left[1 + \exp\left(-\left(\mathbf{c}_{ij}^\top \mathbf{z}\right)b_{ij}\right)\right].$$

In our setting, we have $n = 10$, $m_i = 10$, for all i, and $p = 3$.

A. Convergence rate

In our first experiment, we compare the convergence rate of algorithms that solve the above distributed consensus optimization problem over directed graphs, including ADD-OPT, DEXTRA, [37], Gradient-Push, [26], Directed-Distributed Gradient Descent, [21], and the Weight Balanced-Distributed Gradient Descent, [34]. The network topology is described in Fig. 1, where we apply the weighting strategy from Eq. (3). The step-size used in Gradient-Push, Directed-Distributed Gradient Descent, and Weight Balanced-Distributed Gradient Descent is $\alpha_k = 1/\sqrt{k}$. The constant step-size used in DEXTRA and ADD-OPT is $\alpha = 0.3$. The convergence rates for these algorithms are shown in Fig. 2. It shows that ADD-OPT and DEXTRA have a fast linear convergence rate, while other methods are sub-linear.
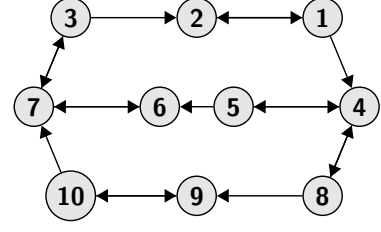

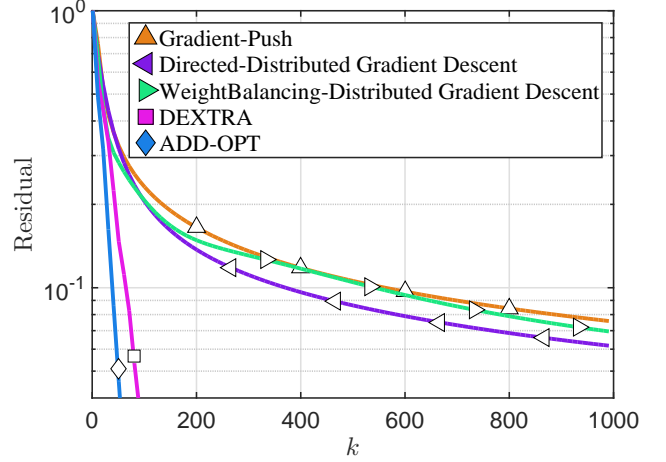Fig. 1: A strongly-connected directed network.

Fig. 2: Convergence rates comparison over directed networks.

B. Step-size range

We now compare ADD-OPT and DEXTRA in terms of their step-size ranges again with the weighting strategy from Eq. (3). It is shown in Fig. 3 that the greatest lower bound of DEXTRA is around $\underline{\alpha} = 0.2$. In contrast, ADD-OPT works for a sufficiently small step-size. In the given setting, we have $\tau = 1.25$, $\epsilon = 1.11$, $y = 1.96$, $y_- = 2.2$, $l = 1$, and $\sigma < 1$; resulting into $\overline{\alpha} = \frac{\sqrt{8.7}}{9.57}$, where we choose c and d to be 1. It can be found in Fig. 4 that the practical upper bound of step-size is much bigger, i.e., $\overline{\alpha} = 1.12$. Since the computation of $\overline{\alpha}$ is related to the global knowledge, e.g., the network topology, and the strong-convexity and Lipschitz-continuity constants, it is preferable to estimate $\overline{\alpha}$. According to Eq. (29), we have that $\overline{\alpha} \cong \sqrt{\frac{s(1-\sigma)^2}{\epsilon yy_-^2(l+s)l^2}}$ given that $\epsilon y(l+s)s(1-\sigma)^2 \gg (\epsilon \tau s)^2$. By estimating $\tau = \epsilon = y = y_- = 1$, $\sigma = 0.9$, and noting that $s \leq l$, we can estimate $\overline{\alpha}$ as $\overline{\alpha} \cong \frac{1}{10l}$.

C. Convergence rate vs. step-sizes

We note that the convergence rate of ADD-OPT is related to the spectral radius of matrix G, i.e., $\rho(G)$, see Eq. (28). Therefore, it is possible to achieve the best convergence rate by picking some α such that the $\rho(G)$ is minimized. In Fig. 5, we show the relationship between the spectral radius, $\rho(G_\alpha)$, of G, and the step-size, α, as well as the residual at the 200-th iteration, $\frac{\|\mathbf{z}_{200} - \mathbf{z}^*\|}{\|\mathbf{z}_0 - \mathbf{z}^*\|}$, and α. We observe that the best convergence rate is achieved when $\alpha = 0.3$, at which $\rho(G)$ is minimized. Fig. 5 also demonstrates our previous theoretical analysis in Lemma 4, where we show that $\rho(G) = 1$,
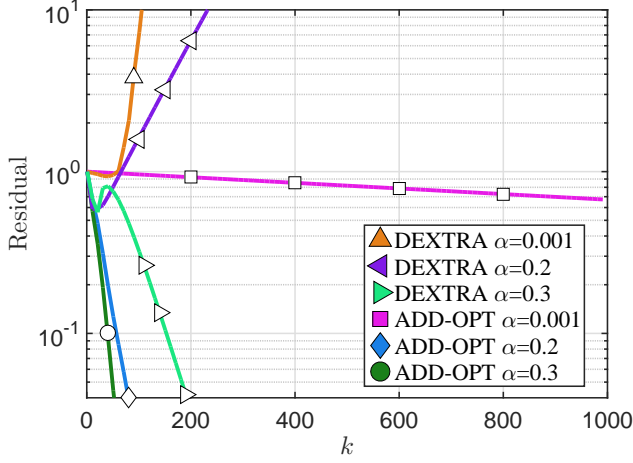
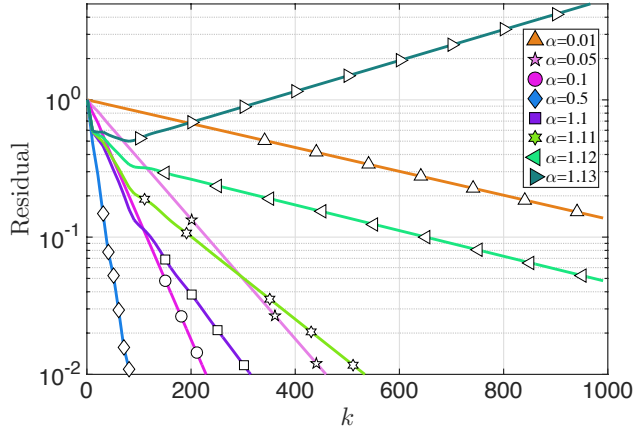Fig. 3: Comparison between ADD-OPT and DEXTRA in terms of step-sizes.

Fig. 4: The range of ADD-OPT 's step-size.

when $\alpha = 0$ or $\alpha = \overline{\alpha}$, and $\rho(G) < 1$ for $\alpha \in (0, \overline{\alpha})$. We further note that $\rho(G_\alpha) < 1$, when α lies approximately in $(0, 0.3)$, which is our theoretical bound of the step-size.
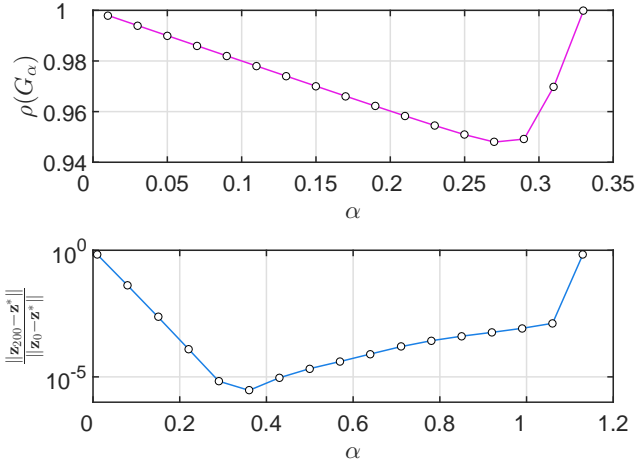

Fig. 5: Spectral radius, $\rho(G_\alpha)$ and the residual at the 200th iteration versus α.

D. Convergence rate vs graph sparsity

In our last experiment, we observe how does the convergence rate change as a function of the sparsity of the directed graph. We consider three strongly-connected directed graphs as shown in Fig. 6. It can be observed that the residuals decrease faster as the number of edges increases, from $\mathcal{G}_a$ to $\mathcal{G}_b$ to $\mathcal{G}_c$, see Fig. 7. This indicates faster convergence when there are more communication channels available for information exchange.
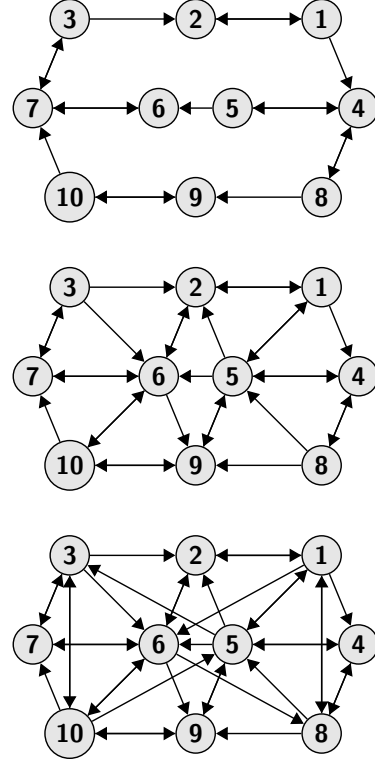

Fig. 6: Three examples of strongly-connected directed graphs.

VII. CONCLUSIONS

In this paper, we focus on solving the distributed optimization problem over directed graphs. The proposed algorithm, termed ADD-OPT (Accelerated Distributed Directed Optimization), can be viewed as an improvement of our recent work, DEXTRA. The proposed algorithm, ADD-OPT, achieves the best known rate of convergence for this class of problems, $O(\mu^k), 0 < \mu < 1$, given that the objective functions are strongly-convex with globally Lipschitz-continuous gradients, where k is the number of iterations. Moreover, ADD-OPT supports a wider and more realistic range of step-sizes in contrast to the existing work. In particular, we show that ADD-OPT converges for arbitrarily small (positive) step-sizes. Simulations further illustrate our results.

REFERENCES

[1] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundation*
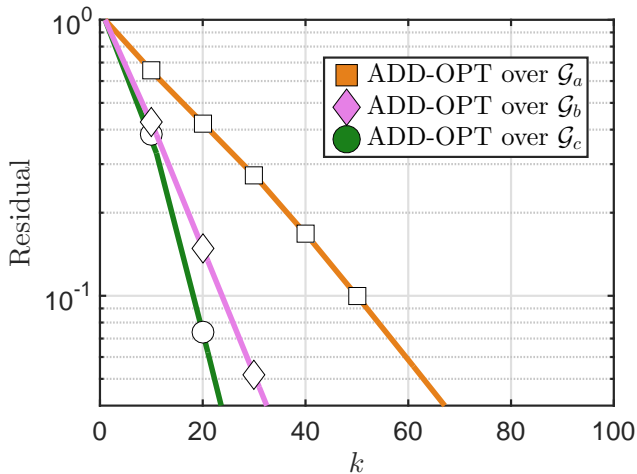
Fig. 7: The range of ADD-OPT 's step-size.

and Trends in Maching Learning, vol. 3, no. 1, pp. 1–122, Jan. 2011.

[2] V. Cevher, S. Becker, and M. Schmidt, "Convex optimization for big data: Scalable, randomized, and parallel algorithms for big data analytics," *IEEE Signal Processing Magazine*, vol. 31, no. 5, pp. 32–43, 2014.

[3] I. Necoara and J. A. K. Suykens, "Application of a smoothing technique to decomposition in convex optimization," *IEEE Transactions on Automatic Control*, vol. 53, no. 11, pp. 2674–2679, Dec. 2008.

[4] G. Mateos, J. A. Bazerque, and G. B. Giannakis, "Distributed sparse linear regression," *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5262–5276, Oct. 2010.

[5] J. A. Bazerque and G. B. Giannakis, "Distributed spectrum sensing for cognitive radio networks by exploiting sparsity," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1847–1862, March 2010.

[6] M. Rabbat and R. Nowak, "Distributed optimization in sensor networks," in *3rd International Symposium on Information Processing in Sensor Networks*, Berkeley, CA, Apr. 2004, pp. 20–27.

[7] U. A. Khan, S. Kar, and J. M. F. Moura, "Diland: An algorithm for distributed sensor localization with noisy distance measurements," *IEEE Transactions on Signal Processing*, vol. 58, no. 3, pp. 1940–1947, Mar. 2010.

[8] C. L and L. Li, "A distributed multiple dimensional qos constrained resource scheduling optimization policy in computational grid," *Journal of Computer and System Sciences*, vol. 72, no. 4, pp. 706 – 726, 2006.

[9] G. Neglia, G. Reina, and S. Alouf, "Distributed gradient optimization for epidemic routing: A preliminary evaluation," in *2nd IFIP in IEEE Wireless Days*, Paris, Dec. 2009, pp. 1–6.

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

[11] A. Nedic, A. Ozdaglar, and P. A. Parrilo, "Constrained consensus and optimization in multi-agent networks," *IEEE Transactions on Automatic Control*, vol. 55, no. 4, pp. 922–938, Apr. 2010.

[12] I. Lobel and A. Ozdaglar, "Distributed subgradient methods for convex optimization over random networks," *IEEE Transactions on Automatic Control*, vol. 56, no. 6, pp. 1291–1306, Jun. 2011.

[13] S. S. Ram, A. Nedic, and V. V. Veeravalli, "Distributed stochastic subgradient projection algorithms for convex optimization," *Journal of Optimization Theory and Applications*, vol. 147, no. 3, pp. 516–545, 2010.

[14] K. Yuan, Q. Ling, and W. Yin, "On the convergence of decentralized gradient descent," *arXiv preprint arXiv:1310.7063*, 2013.

[15] D. Jakovetic, J. Xavier, and J. M. F. Moura, "Fast distributed gradient methods," *IEEE Transactions on Automatic Control*, vol. 59, no. 5, pp. 1131–1146, 2014.

[16] 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.

[17] G. Qu and N. Li, "Harnessing smoothness to accelerate distributed optimization," *arXiv preprint arXiv:1605.07112*, 2016.

[18] J. F. C. Mota, J. M. F. Xavier, P. M. Q. Aguiar, and M. Puschel, "D-ADMM: A communication-efficient distributed algorithm for separable optimization," *IEEE Transactions on Signal Processing*, vol. 61, no. 10, pp. 2718–2723, May 2013.

[19] E. Wei and A. Ozdaglar, "Distributed alternating direction method of multipliers," in *51st IEEE Annual Conference on Decision and Control*, Dec. 2012, pp. 5445–5450.

[20] W. Shi, Q. Ling, K Yuan, G Wu, and W Yin, "On the linear convergence of the admm in decentralized consensus optimization," *IEEE Transactions on Signal Processing*, vol. 62, no. 7, pp. 1750–1761, April 2014.

[21] C. Xi, Q. Wu, and U. A. Khan, "Distributed gradient descent over directed graphs," *arXiv preprint arXiv:1510.02146*, 2015.

[22] 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.

[23] T. Charalambous, M. G. Rabbat, M. Johansson, and C. N. Hadjicostis, "Distributed finite-time computation of digraph parameters: Left-eigenvector, out-degree and spectrum," *IEEE Transactions on Control of Network Systems*, vol. 3, no. 2, pp. 137–148, June 2016.

[24] B. Gharesifard and J. Cortés, "Distributed continuous-time convex optimization on weight-balanced digraphs," *IEEE Transactions on Automatic Control*, vol. 59, no. 3, pp. 781–786, March 2014.

[25] S. S. Kia, J. Cortes, and S. Martinez, "Distributed convex optimization via continuous-time coordination algorithms with discrete-time communication," *Automatica*, vol. 55, pp. 254 – 264, 2015.

[26] A. Nedic and A. Olshevsky, "Distributed optimization

over time-varying directed graphs," *IEEE Transactions on Automatic Control*, vol. 60, no. 3, pp. 601–615, Mar. 2015.

[27] 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.

[28] K. I. Tsianos, *The role of the Network in Distributed Optimization Algorithms: Convergence Rates, Scalability, Communication/Computation Tradeoffs and Communication Delays*, Ph.D. thesis, Dept. Elect. Comp. Eng. McGill University, 2013.

[29] K. I. Tsianos, S. Lawlor, and M. G. Rabbat, "Consensus-based distributed optimization: Practical issues and applications in large-scale machine learning," in *50th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL, USA, Oct. 2012, pp. 1543–1550.

[30] 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.

[31] F. Benezit, V. Blondel, P. Thiran, J. Tsitsiklis, and M. Vetterli, "Weighted gossip: Distributed averaging using non-doubly stochastic matrices," in *IEEE International Symposium on Information Theory*, Jun. 2010, pp. 1753–1757.

[32] C. Xi and U. A. Khan, "Distributed subgradient projection algorithm over directed graphs," *arXiv preprint arXiv:1602.00653*, 2016.

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

[34] A. Makhdoumi and A. Ozdaglar, "Graph balancing for distributed subgradient methods over directed graphs," *to appear in 54th IEEE Annual Conference on Decision and Control*, 2015.

[35] L. Hooi-Tong, "On a class of directed graphswith an application to traffic-flow problems," *Operations Research*, vol. 18, no. 1, pp. 87–94, 1970.

[36] A. Nedic and A. Olshevsky, "Distributed optimization of strongly convex functions on directed time-varying graphs," in *IEEE Global Conference on Signal and Information Processing*, Dec. 2013, pp. 329–332.

[37] C. Xi and U. A. Khan, "On the linear convergence of distributed optimization over directed graphs," *arXiv preprint arXiv:1510.02149*, 2015.

[38] J. Zeng and W. Yin, "Extrapush for convex smooth decentralized optimization over directed networks," *arXiv preprint arXiv:1511.02942*, 2015.

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

[40] R. Bhatia, *Matrix analysis*, vol. 169, Springer Science & Business Media, 2013.

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

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

[43] B. Polyak, *Introduction to optimization*, Optimization Software, 1987.

Chenguang Xi received his B.S. degree in Microelectronics from Shanghai JiaoTong University, China, in 2010, M.S. and Ph.D. degrees in Electrical and Computer Engineering from Tufts University, in 2012 and 2016, respectively. His research interests include distributed optimization, tensor analysis, and source localization.

Ran Xin received his B.S. degree in Mathematics and Applied Mathematics from Xiamen University, China, in 2016. His research interests include distributed optimization and control.

Usman A. Khan received his B.S. degree (with honors) in EE from University of Engineering and Technology, Lahore-Pakistan, in 2002, M.S. degree in ECE from University of Wisconsin-Madison in 2004, and Ph.D. degree in ECE from Carnegie Mellon University in 2009. Currently, he is an Assistant Professor with the ECE Department at Tufts University. He received the NSF Career award in Jan. 2014 and is an IEEE Senior Member since Feb. 2014. His research interests lie in efficient operation and planning of complex infrastructures and include statistical signal processing, networked control and estimation, and distributed algorithms. Dr. Khan is on the editorial board of IEEE Transactions on Smart Grid and an associate member of Sensor Array and Multichannel Technical Committee with the IEEE Signal Processing Society. He has served on the Technical Program Committees of several IEEE conferences and has organized and chaired several IEEE workshops and sessions. His graduate students have won multiple Best Student Paper awards. His work was presented as Keynote speech at BiOS SPIE Photonics West–Nanoscale Imaging, Sensing, and Actuation for Biomedical Applications IX.