AB/Push-Pull Method for Distributed Optimization in Time-Varying Directed Networks

Angelia Nedić[a] and Duong Thuy Anh Nguyen[a] and Duong Tung Nguyen[a]

[a]School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, AZ, United States

ARTICLE HISTORY
Compiled September 14, 2022

ABSTRACT
In this paper, we study the distributed optimization problem for a system of agents embedded in time-varying directed communication networks. Each agent has its own cost function and agents cooperate to determine the global decision that minimizes the summation of all individual cost functions. We consider the so-called push-pull gradient-based algorithm (termed as *AB*/Push-Pull) which employs both row- and column-stochastic weights simultaneously to track the optimal decision and the gradient of the global cost while ensuring consensus and optimality. We show that the algorithm converges linearly to the optimal solution over a time-varying directed network for a constant stepsize when the agent's cost function is smooth and strongly convex. The linear convergence of the method has been shown in Saadatniaki et al. (2020), where the multi-step consensus contraction parameters for row- and column-stochastic mixing matrices are not directly related to the underlying graph structure, and the explicit range for the stepsize value is not provided. With respect to Saadatniaki et al. (2020), the novelty of this work is twofold: (1) we establish the one-step consensus contraction for both row- and column-stochastic mixing matrices with the contraction parameters given explicitly in terms of the graph diameter and other graph properties; and (2) we provide explicit upper bounds for the stepsize value in terms of the properties of the cost functions, the mixing matrices, and the graph connectivity structure.

KEYWORDS
Distributed optimization; gradient tracking; time-varying graphs; directed graphs

1. Introduction

We consider a system of n agents embedded in a communication network with the goal to collaboratively solve the following minimization problem:

$$\min_{x\in\mathbb{R}^p} \; f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x), \tag{1}$$

where each function $f_i : \mathbb{R}^p \to \mathbb{R}$ represents the cost function of agent i, is strongly convex and known by agent i only. The strong convexity condition implies that problem (1) has a unique optimal solution. The agents want to determine the optimal

CONTACT Angelia Nedić. Email: Angelia.Nedich@asu.edu. This work has been partially supported by the Office of Naval Research award N00014-21-1-2242

solution by performing local computations and limited information exchange with their local neighbors in the communication network. Decentralized and collaborative approach is particularly appealing in large-scale, multi-agent systems with privacy concerns and limited computation, communication, or storage capabilities. In these scenarios, the data is collected and/or stored in a distributed manner, thus, computing tasks are distributed over all the agents and information exchange occurs only between the agents with direct communication links. Such problems appear in many engineering and scientific applications for example in wireless sensor networks [19], distributed sensing [1], trajectory optimization for formation control of vehicles [27], large-scale machine learning [29], and cooperative multi-agent systems [20].

Distributed optimization of the sum of convex functions has been of considerable interest and many algorithms have been proposed including gradient-based methods [7, 8, 21, 26, 35], dual averaging methods [2], ADMM [25], and Newton methods [6, 30]. Early works have often assumed that the underlying network is undirected (see literature review in [5]) and most commonly require doubly stochastic or weight-balanced [4] mixing matrices. Reference [24] uses a gradient difference structure in the algorithm to provide the first-order method that achieves a geometric convergence with the requirement of symmetric weights. Based on the ADMM approach, the work in [25] demonstrates a linear convergence while the Nesterov's acceleration method in reference [12] obtains convergence times that scale linearly in the number of agents. Reference [23] investigates decentralized algorithms that take advantage of proximal operations for the non-smooth terms. In [15, 16], stochastic variants of distributed methods have been considered for asynchronous computations.

In many scenarios, agents communications are directed such as, for example, due to broadcasting at different power levels, thus resulting in communications that correspond to directed graphs. To cope with directed graphs, reference [28] introduces a subgradient-push algorithm to achieve consensus among the agents on an optimal point. The work in [9] further studies the push-sum technique for time-varying directed graphs with a convergence rate of $O(\ln t/\sqrt{t})$ for diminishing stepsizes. Aiming to improve the convergence rate, algorithms ADD-OPT [33] and Push-DIGing [10] incorporates the push-sum protocol with gradient estimation approach, and show geometric convergence for a sufficiently small step-size. The implementation of these methods require the knowledge of agents' out-degree in order to construct a column-stochastic weight matrix, which is later removed in [32] and in FROST method [36].

The aforementioned push-sum based works use an independent algorithm to asymptotically compute the right or left eigenvectors of the weight matrix, corresponding to the eigenvalue of 1. Thus, the resulting algorithms are nonlinear and involve additional computation among agents. Unlike the push-sum protocol, the alternate AB/Push-Pull methods introduced in [17, 35] use a row-stochastic matrix and a column-stochastic matrix simultaneously to achieve a linear convergence. Recent work in [14] further addresses the challenge of noisy information exchange and shows linear convergence (in expectation) to a neighborhood of the optimum exponentially fast, under a constant stepsize. The analysis of AB/Push-Pull (with stochastic gradients) was shown in [34]. A variant of the method, where the stepsize α is agent dependent, has been analyzed in [17] for the case of a static graph. All the aforementioned work on the AB/Push-Pull methods is for a static directed graph. The AB/Push-Pull method for time-varying directed graphs has been studied in [22], where a linear convergence is shown for the case when the global objective function is smooth and strongly-convex, and the underlying time-varying graphs have bounded connectivity. In order to facilitate privacy design, the recent work in [31] proposes to tailor gradient methods for differentially-private

distributed optimization. The work in [3] provides a general gradient-tracking based privacy-preserving algorithm with added randomness in optimization parameters and shows that the added randomness has no influence on the accuracy of optimization.

In this paper, we consider AB/Push-Pull algorithm where the agent communications are given by a sequence of time-varying directed graphs. At every time k, the agents' updates are described by two non-negative matrices that are compliant with the connectivity structure of the graph: a row-stochastic matrix for the mixing of the decision variables *(pull-step)* and a column-stochastic matrix for tracking the average gradients *(push-step)*. We prove that the method converges to the optimal solution geometrically fast, provided that the stepsize is small enough and the agents' objective functions are smooth enough. Moreover, we provide an explicit upper bound for the stepsize range and characterize the convergence rate in terms of the problem parameters, algorithms' parameters (weight matrices), and the underlying graphs' connectivity structures.

A key difficulty in the analysis is imposed by the time-varying nature of the mixing matrices. Our analysis makes use of time-varying weighted averages and time-varying weighted norms, where the weights are defined in terms of stochastic vector sequences associated with the mixing matrix sequences. This allows us to establish consensus contractions per each update step for both row- and column-stochastic mixing matrices. This is unlike the work in [22] that considers the AB/Push-Pull method over time-varying graphs, where the analysis makes use of the Euclidean norms – at the expense of relying on a multi-step consensus contraction, even when every underlying graph is strongly connected. Moreover, through the use of time-varying weighted norms and the relations of the weight matrices with the underlying graphs, we provide explicit upper bounds for the stepsize range in terms of properties of the mixing matrices and the graphs' connectivity structure. This is in sharp contrast with [22] where no explicit range is provided. Also, our analysis in this paper is in contrast with [17, 34] where the stepsize range is given in terms of the singular values of the weight matrices, which are neither explicitly capturing the structure of the matrices nor the underlying graph connectivity structure.

The structure of this paper is as follows. We first provide notation, introduce our algorithm and state basic assumptions in Section 2. We present some basic results in Section 3. We establish the convergence properties of the algorithm in Section 4 and Section 5, and we conclude with some remarks in Section 7.

2. Notation and Terminology

Throughout the paper, all vectors are viewed as column-vectors unless stated otherwise. We use $\langle \cdot, \cdot \rangle$ to denote the inner product, and $\| \cdot \|$ to denote the standard Euclidean norm. We write $\mathbf{1}$ to denote the vector with all entries equal to 1, and $\mathbb{I}$ to denote the identity matrix. The i-th entry of a vector u is denoted by u_i, while it is denoted by $[u_k]_i$ for a time-varying vector u_k. Given a vector v, we use $\min(v)$ and $\max(v)$ to denote the smallest and the largest entry of v, respectively, i.e., $\min(v) = \min_i v_i$ and $\max(v) = \max_i v_i$. A vector is said to be a stochastic vector if its entries are nonnegative and sum to 1.

To denote the ij-th entry of a matrix A, we write A_{ij}, and we write $[A_k]_{ij}$ when the matrix is time-dependent. For any two matrices A and B of the same dimension, we write $A \leq B$ to denote that $A_{ij} \leq B_{ij}$ for all i and j. A matrix is said to be nonnegative if all its entries are nonnegative. For a nonnegative matrix, we use $\min(A^+)$ to denote the smallest positive entry of A, i.e., $\min(A^+) = \min_{\{ij:A_{ij}>0\}} A_{ij}$. A nonnegative

matrix is said to be row-stochastic if each row entries sum to 1, and it is said to be column-stochastic if each column entries sum to 1. In particular, if $A \in \mathbb{R}^{n\times n}$ is row-stochastic and $B \in \mathbb{R}^{n\times n}$ is column stochastic, then $A\mathbf{1} = \mathbf{1}$ and $\mathbf{1}^T B = \mathbf{1}^T$.

Given a vector $\mathbf{a} \in \mathbb{R}^n$ with positive entries $a_1, \ldots, a_n$, the $\mathbf{a}$-weighted norm can be induced in the vector space $\mathbb{R}^p \times \cdots \times \mathbb{R}^p$ (consisting of n copies of $\mathbb{R}^p$), as follows:

$$\|\mathbf{x}\|_{\mathbf{a}} = \sqrt{\sum_{i=1}^{n} a_i \|x_i\|^2} \qquad \text{for } \mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^p \times \cdots \times \mathbb{R}^p.$$

When $\mathbf{a} = \mathbf{1}$, we simply write $\|\mathbf{x}\|$. We also write $\|\mathbf{x}\|_{\mathbf{a}^{-1}}$ to denote the norm induced by the vector with entries $1/a_i$, i.e., $\|\mathbf{x}\|_{\mathbf{a}^{-1}} = \sqrt{\sum_{i=1}^{n} \frac{\|x_i\|^2}{a_i}}$. The following relations, which can be proved by using Hölder's inequality, will be useful in our analysis:

$$\|\mathbf{x}\| \leq \sqrt{\tfrac{1}{\min(\mathbf{a})}}\, \|\mathbf{x}\|_{\mathbf{a}} \qquad \text{for all } \mathbf{x} \in \mathbb{R}^p \times \cdots \times \mathbb{R}^p \text{ and } \mathbf{a} > \mathbf{0}, \tag{2a}$$

$$\|\mathbf{x}\| \leq \|\mathbf{x}\|_{\mathbf{a}^{-1}} \qquad \text{for all } \mathbf{x} \in \mathbb{R}^p \times \cdots \times \mathbb{R}^p \text{ and } \mathbf{a} > 0 \text{ satisfying } \langle \mathbf{a}, \mathbf{1} \rangle = 1. \tag{2b}$$

We let $[n] = \{1, \ldots, n\}$ for an integer $n \geq 1$. A directed graph $\mathbb{G} = ([n], \mathcal{E})$ is specified by the edge set $\mathcal{E} \subseteq [n] \times [n]$ of ordered pairs of nodes. Given a directed graph $\mathbb{G} = ([n], \mathcal{E})$, the sets $\mathcal{N}_i^{\text{out}}$ and $\mathcal{N}_i^{\text{in}}$ denote the out-neighbors and the in-neighbors of a node i, i.e., $\mathcal{N}_i^{\text{out}} = \{j \mid (i,j) \in \mathcal{E}\}$ and $\mathcal{N}_i^{\text{in}} = \{j \mid (j,i) \in \mathcal{E}\}$.

We say that a directed graph is *strongly connected* if there is a directed path from any node to all other nodes in the graph. Given a directed path, the length of the path is the number of edges in the path. We use the following definitions:

Definition 2.1 (Graph Diameter)**.** The diameter of a strongly connected directed graph $\mathbb{G}$, denoted by $\mathsf{D}(\mathbb{G})$, is the length of the longest path in the collection of all shortest directed paths connecting all ordered pairs of distinct nodes in $\mathbb{G}$.

Let $\mathbf{p}_{jl}$ denote a *shortest directed path from node j to node l*, where $j \neq l$. A collection $\mathcal{P}$ of directed paths in $\mathbb{G}$ is a shortest-path graph covering if $\mathbf{p}_{jl} \in \mathcal{P}$ and $\mathbf{p}_{lj} \in \mathcal{P}$ for any two nodes $j,\ l \in [n]$, $j \neq l$. The *utility of the edge* (j,l) with respect to the covering $\mathcal{P}$ is the number of shortest paths in $\mathcal{P}$ that pass through the edge (j,l). Define $\mathsf{K}(\mathcal{P})$ as the maximum edge-utility in $\mathcal{P}$ taken over all edges in the graph, i.e., $\mathsf{K}(\mathcal{P}) = \max\limits_{(j,l)\in\mathcal{E}} \sum\limits_{\mathbf{p}\in\mathcal{P}} \chi_{\{(j,l)\in\mathbf{p}\}}$, where $\chi_{\{(j,l)\in\mathbf{p}\}}$ is the indicator function taking value 1 when $(j,l) \in \mathbf{p}$ and, otherwise, taking value 0. Denote by $\mathcal{S}(\mathbb{G})$ the collection of all possible shortest-path coverings of the graph $\mathbb{G}$, we have the following definition.

Definition 2.2 (Maximal Edge-Utility)**.** For a strongly connected directed graph $\mathbb{G} = ([n], \mathcal{E})$, the maximal edge-utility is the maximum value of $\mathsf{K}(\mathcal{P})$ taken over all possible shortest-path coverings $\mathcal{P} \in \mathcal{S}(\mathbb{G})$, i.e., $\mathsf{K}(\mathbb{G}) = \max_{\mathcal{P}\in\mathcal{S}(\mathbb{G})} \mathsf{K}(\mathcal{P})$.

2.1. AB/Push-Pull Method and Assumptions

We consider a system with n agents, and let each agent $i \in \{1, 2, \ldots, n\}$ have a local copy $x_i \in \mathbb{R}^p$ of the decision variable and a direction $y_i \in \mathbb{R}^p$ which is an estimate of a "global update direction". These variables are maintained and updated over time and at iteration k, they are denoted by x_i^k and y_i^k, respectively. We present a dis-

tributed algorithm, termed AB/Push-Pull algorithm to fairly capture independent and simultaneous developments of two closely related methods, namely the *Push-Pull* method of [17] and the method proposed in [35]. We consider the AB/Push-Pull gradient method over a sequence $\{\mathbb{G}_k\}$ of directed graphs, where the agents communicate over a graph $\mathbb{G}_k$ at the round k of updates. At every time k, the agents updates are described by two non-negative matrices A_k and B_k that are compliant with the connectivity structure of the graph $\mathbb{G}_k$, i.e.,

$$[A_k]_{ij} > 0 \quad \text{for all } j \in \mathcal{N}_{ik}^{\text{in}} \cup \{i\}, \qquad [A_k]_{ij} = 0 \quad \text{for all } j \notin \mathcal{N}_{ik}^{\text{in}} \cup \{i\}, \tag{3a}$$

$$[B_k]_{ji} > 0 \quad \text{for all } j \in \mathcal{N}_{ik}^{\text{out}} \cup \{i\}, \qquad [B_k]_{ji} = 0 \quad \text{for all } j \notin \mathcal{N}_{ik}^{\text{out}} \cup \{i\}. \tag{3b}$$

Moreover, each matrix A_k is row-stochastic and each matrix B_k is column-stochastic, i.e., $A_k\mathbf{1} = \mathbf{1}$ and $\mathbf{1}^T B_k = \mathbf{1}^T$ for all $k \geq 0$. The method works as follows: at time k, every agent i sends its vector x_i^k and a scaled direction $[B_k]_{ji}y_i^k$ to its out-neighbors $j \in \mathcal{N}_{ik}^{\text{out}}$, while it keeps $[B_k]_{ii}y_i^k$ for its own update.

Upon the information exchange step, every agent i updates as follows: for all $k \geq 0$,

$$x_i^{k+1} = \sum_{j \in \mathcal{N}_{ik}^{\text{in}}} [A_k]_{ij} x_j^k - \alpha y_i^k, \tag{4a}$$

$$y_i^{k+1} = \sum_{j \in \mathcal{N}_{ik}^{\text{in}}} [B_k]_{ij} y_j^k + \nabla f_i(x_i^{k+1}) - \nabla f_i(x_i^k), \tag{4b}$$

where $\alpha > 0$ is a stepsize. In this method, the agent i decides on the entries of A_k in the ith row (for the in-neighbors $j \in \mathcal{N}_{ik}^{\text{in}}$), while the value $[B_k]_{ij}$ is selected by agent $j \in \mathcal{N}_{ik}^{\text{in}}$. Each agent i initializes the updates with an arbitrary vector x_i^0 and with $y_i^0 = \nabla f_i(x_i^0)$, which does not require any coordination among agents. The update step using the mixing matrix A_k is viewed as a *pull-step*, while the step utilizing the matrix B_k is viewed as a *push-step* as it is reminiscent of the push-sum consensus method.

When the matrices A_k and B_k are compatible with the underlying graph $\mathbb{G}_k$ (see (3a) and (3b)), we can re-write the method (4) as follows: for all $i \in [n]$ and all $k \geq 0$,

$$x_i^{k+1} = \sum_{j=1}^{n} [A_k]_{ij} x_j^k - \alpha y_i^k, \tag{5a}$$

$$y_i^{k+1} = \sum_{j=1}^{n} [B_k]_{ij} y_j^k + \nabla f_i(x_i^{k+1}) - \nabla f_i(x_i^k), \tag{5b}$$

$$\text{where } x_i^0 \in \mathbb{R}^p \text{ is arbitrary and } y_i^0 = \nabla f_i(x_i^0). \tag{5c}$$

We analyze the convergence properties of the method under the following assumptions:

Assumption 1 (Strongly Connected Graphs)**.** For each k, the directed graph $\mathbb{G}_k = ([n], \mathcal{E}_k)$ is strongly connected.

Assumption 2 (Graph Compatible A_k)**.** For each k, the matrix A_k is row-stochastic and compatible with the graph $\mathbb{G}_k$ in the sense of relation (3a). Moreover, there exists a scalar $a > 0$ such that $\min(A_k^+) \geq a$ for all $k \geq 0$.

Assumption 3 (Graph Compatible B_k)**.** For each k, the matrix B_k is column-stochastic and compatible with the graph $\mathbb{G}_k$ in the sense of relation (3b). Moreover, there exists a scalar $b > 0$ such that $\min(B_k^+) \geq b$ for all $k \geq 0$.

Assumption 4 (Lipschitz gradient)**.** Each f_i is continuously differentiable and has L-Lipschitz continuous gradients, i.e., for some $L > 0$,

$$\|\nabla f_i(x) - \nabla f_i(u)\| \leq L\|x - u\|, \qquad \text{for all } x, u \in \mathbb{R}^p.$$

Assumption 5 (Strong convexity)**.** The average-sum function $f = \frac{1}{n}\sum_{i=1}^n f_i$ is μ-strongly convex, i.e., for some $\mu > 0$,

$$\langle \nabla f(x) - \nabla f(u), x - u\rangle \geq \mu\|x - u\|^2 \qquad \text{for all } x, u \in \mathbb{R}^p.$$

3. Basic Results

3.1. Linear Combinations and Graphs

Certain contractive properties of the iterates produced by the method are inherited from the use of the mixing terms $\sum_{j=1}^n [A_k]_{ij} x_j^k$ and $\sum_{j=1}^n [B_k]_{ij} y_j^k$, and the fact that the matrices A_k and B_k are compliant with a directed strongly connected graph $\mathbb{G}_k$. The following results will help us capture these contractive properties.

For a collection $\{u_i,\, i \in [n]\} \subset \mathbb{R}^p$ of vectors and a collection $\{\gamma_i,\, i \in [n]\} \subset \mathbb{R}$ of scalars, we have the following relations (see Lemma 1 and Corollary 1 of [11]):

$$\left\|\sum_{i=1}^n \gamma_i u_i\right\|^2 = \left(\sum_{j=1}^n \gamma_j\right)\sum_{i=1}^n \gamma_i\|u_i\|^2 - \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n \gamma_i\gamma_j\|u_i - u_j\|^2. \tag{6}$$

Moreover, if $\sum_{i=1}^n \gamma_i = 1$ holds, then we have

$$\frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n \gamma_i\gamma_j\|u_i - u_j\|^2 = \sum_{i=1}^n \gamma_i \left\|u_i - \left(\sum_{\ell=1}^n \gamma_\ell u_\ell\right)\right\|^2, \tag{7a}$$

$$\left\|\sum_{i=1}^n \gamma_i u_i - u\right\|^2 = \sum_{i=1}^n \gamma_i\|u_i - u\|^2 - \sum_{i=1}^n \gamma_i \left\|u_i - \left(\sum_{\ell=1}^n \gamma_\ell u_\ell\right)\right\|^2, \quad \text{for all } u \in \mathbb{R}^p. \tag{7b}$$

We also make use of the following result.

Lemma 3.1 ([11], Lemma 2)**.** *Let $\mathbb{G} = ([n], \mathcal{E})$ be a strongly connected directed graph, where a vector x_i is associated with node i for all $i \in [n]$. We then have*

$$\sum_{(j,\ell)\in\mathcal{E}} \|x_j - x_\ell\|^2 \geq \frac{1}{\mathsf{D}(\mathbb{G})\mathsf{K}(\mathbb{G})}\sum_{j=1}^n\sum_{\ell=j+1}^n \|x_j - x_\ell\|^2,$$

where $\mathsf{D}(\mathbb{G})$ is the diameter of the graph $\mathbb{G}$ and $\mathsf{K}(\mathbb{G})$ is the maximal edge-utility in the graph (see Definitions 2.1 and 2.2).

3.2. Implications of Stochastic Nature of Matrices A_k and B_k

The column stochastic property of the matrices B_k ensures that the sum of the y-iterates of the method (5), at any time k, is equal to the sum of the gradients $\nabla f_i(x_i^k)$, as seen in the following lemma.

Lemma 3.2. *Consider the method in (5), and assume that each B_k is column-stochastic. Then, we have $\sum_{i=1}^n y_i^k = \sum_{i=1}^n \nabla f_i(x_i^k)$ for all $k \geq 0$.*

Proof. The proof is by the mathematical induction on k. ∎

Lemma 3.3 ([11], Lemma 3)**.** *Let Assumption 1 hold, and let $\{A_k\}$ be a matrix sequence satisfying Assumption 2. Then, there exists a sequence $\{\phi_k\}$ of stochastic vectors such that*

$$\phi_{k+1}^T A_s = \phi_k^T \qquad \text{for all } k \geq 0, \tag{8}$$

where the entries of each ϕ_k are positive and have a uniform lower bound, i.e.,

$$[\phi_k]_i \geq \frac{a^n}{n} \qquad \text{for all } i \in [n],$$

with $a \in (0,1)$ being the lower bound on the positive entries of the matrices A_k.

For the matrices B_k, we define the stochastic vector sequence $\{\pi_k\}$ as follows:

$$\pi_{k+1} = B_k \pi_k, \qquad \text{initialized with } \ \pi_0 = \frac{1}{n}\mathbf{1}. \tag{9}$$

We examine the sequence $\{\pi_k\}$ in the following lemma.

Lemma 3.4. *Let Assumption 1 hold and let the matrix sequence $\{B_k\}$ satisfy Assumption 3. Then, the vectors π_k generated by (9) are stochastic vectors such that*

$$[\pi_k]_i \geq \frac{b^n}{n} \qquad \text{for all } i \in [n] \text{ and } k \geq 0,$$

where $b \in (0,1)$ is the lower bound on the positive entries of the matrices B_k.

Proof. We prove that each π_k is stochastic by using the mathematical induction on k. For $k = 0$, the vector $\pi_0 = \frac{1}{n}\mathbf{1}$ is stochastic. Suppose now the vector π_k is stochastic. Choose any index $i \in [n]$ and consider the entry $[\pi_{k+1}]_i$. By the definition of π_{k+1} in (9), since the entries in B_k and π_k are nonnegative, we have $[\pi_{k+1}]_i = \sum_{j=1}^n [B_k]_{ij}[\pi_k]_j \geq 0$. Furthermore, by summing the entries of π_{k+1}, and using the facts that B_k is column stochastic and π_k is a stochastic vector, we obtain $\mathbf{1}^T\pi_{k+1} = \mathbf{1}^T B_k \pi_k = \mathbf{1}^T\pi_k = 1$. Thus, π_{k+1} is a stochastic vector.

To prove the lower bound result, we consider separately the case for $k = 0, \ldots, n-1$ and the case $k \geq n$. The bound is obviously valid for $k = 0$, since $\pi_0 = \frac{1}{n}\mathbf{1}$. Let k be such that $1 \leq k \leq n-1$. By the definition of π_k, we have

$$\pi_k = B_{k-1} \cdots B_0 \pi_0 = \frac{1}{n} B_{k-1} \cdots B_0 \mathbf{1}.$$

Hence, it follows that

$$[\pi_k]_i = \frac{1}{n}[B_{k-1}\cdots B_0\mathbf{1}]_i = \frac{1}{n}\sum_{j=1}^{n}[B_{k-1}\cdots B_0]_{ij} \geq \frac{1}{n}[B_{k-1}\cdots B_0]_{ii} \geq \frac{b^k}{n},$$

where the last inequality follows from $[B_{k-1}\cdots B_0]_{ii} \geq b^k$, which is valid since all matrices B_k have positive diagonals with diagonal entries larger than or equal to b (see Assumption 3). Since $k < n$, it follows that

$$[\pi_k]_i \geq \frac{b^k}{n} > \frac{b^n}{n} \qquad \text{for all } k = 1, \ldots, n-1.$$

Now, consider the case $k \geq n$. Using the definition of π_k, we obtain

$$\pi_k = B_{k-1}\cdots B_{k-n}\pi_{k-n}.$$

We note that the matrix $[B_{k-1}\cdots B_{k-n}]$ has all entries positive as it represents directed paths among the nodes in the composition of the strongly connected graphs $\mathbb{G}_{k-1}, \ldots, \mathbb{G}_{k-n}$. Moreover, every entry of $[B_{k-1}\cdots B_{k-n}]$ is at least as large as b^n, i.e.,

$$[B_{k-1}\cdots B_{k-n}]_{ij} \geq b^n \qquad \text{for all } i, j \in [n],$$

which follows by Assumption 3 ensuring that each B_t has positive entries on links in the graph $\mathbb{G}_t$, which are at least large as b. Hence, it follows that

$$[\pi_k]_i = \sum_{j=1}^{n}[B_{k-1}\cdots B_{k-n}]_{ij}[\pi_{k-n}]_j \geq \sum_{j=1}^{n} b^n[\pi_{k-n}]_j = b^n > \frac{b^n}{n},$$

where the last equality holds since π_s is a stochastic vector for all s. ■

3.3. Contractive Property of Gradient Method

Lemma 3.5. *For a μ-strongly convex function F with L-Lipschitz continuous gradients, at the point $x^* = \operatorname{argmin}_x F(x)$, we have*

$$\|x - x^* - \alpha\nabla F(x)\| \leq q(\alpha)\|x - x^*\| \qquad \textit{for all } x \textit{ and for all } \alpha \textit{ with } 0 < \alpha < 2L^{-1},$$

where $q(\alpha) = \max\{|1-\alpha\mu|, |1-\alpha L|\} < 1$.

The proof of Lemma 3.5 can be found within the proof of Theorem 3 of Chapter 1 in [13] for a twice continuously differentiable function. The result has been extended in [18] (see Lemma 10 therein) to a more general case of a differentiable function.

4. Convergence Analysis

In this section, we specify and analyze the behavior of three quantities that are critical components of the convergence proof of the method: the distance of a suitably defined weighted average $\hat{x}^k$ from the solution x^* of problem (1), a weighted dispersion of the

iterates x_i^k from the weighted average $\hat{x}^k$, and a weighted dispersion of the agents' directions y_i^k from the sum $\sum_{i=1}^n y_i^k$.

4.1. Weighted Averages of Agents' x-variables

We define $\hat{x}^k$ to be the ϕ_k-weighted averages of the iterates x_i^k produced by the AB/Push-Pull method (5), i.e.,

$$\hat{x}^k = \sum_{i=1}^n [\phi_k]_i x_i^k \qquad \text{for all } k \geq 0, \tag{10}$$

where $\{\phi_k\}$ is the sequence of stochastic vectors satisfying $\phi_{k+1}^T A_k = \phi_k^T$ (see Lemma 3.3). In the next proposition, we establish a recursion relation for $\hat{x}^k$, and a relation for their distance from the optimal solution x^* of problem (1).

Proposition 4.1. *Let Assumptions 2-5 hold. Then, the following statements are valid:*

(a) The weighted average sequence $\{\hat{x}^k\}$ defined in (10) satisfies,

$$\hat{x}^{k+1} = \hat{x}^k - \alpha \sum_{i=1}^n [\phi_{k+1}]_i y_i^k \qquad \text{for all } k \geq 0. \tag{11}$$

(b) Let the stepsize α in method (5) be such that $0 < \alpha < \frac{2}{nL}$, where L is the gradient Lipschitz constant from Assumption 4. Then, we have for all $k \geq 0$,

$$\|\hat{x}^{k+1} - x^*\| \leq q_k(\alpha)\|\hat{x}^k - x^*\| + \alpha L \sqrt{\frac{n}{\min(\phi_k)}} D(\mathbf{x}^k, \phi_k) + \alpha S(\mathbf{y}^k, \pi_k),$$

where $q_k(\alpha) = \max\{|1 - \alpha n \min(\pi_k)\mu|, |1 - \alpha n \min(\pi_k)L|\} < 1$.

Proof. (a) By the definition of $\hat{x}^{k+1}$ and the x-update relation given in (5a), we have

$$\hat{x}^{k+1} = \sum_{i=1}^n [\phi_{k+1}]_i x_i^{k+1} = \sum_{i=1}^n [\phi_{k+1}]_i \sum_{j=1}^n [A_k]_{ij} x_j^k - \alpha \sum_{i=1}^n [\phi_{k+1}]_i y_i^k.$$

For the double-sum term, it follows that

$$\sum_{i=1}^n [\phi_{k+1}]_i \sum_{j=1}^n [A_k]_{ij} x_j^k = \sum_{j=1}^n \left(\sum_{i=1}^n [\phi_{k+1}]_i [A_k]_{ij} \right) x_j^k = \sum_{j=1}^n [\phi_k]_j x_j^k = \hat{x}^k,$$

where the second equality follows by $\phi_{k+1}^T A_k = \phi_k^T$ (see Lemma 3.3), and the last equality uses the definition of $\hat{x}^k$, thus, establishes the desired relation in part (a). (b) Under Assumption 5, the unique minimizer x^* of $f(x)$ over $x \in \mathbb{R}^p$ exists. By subtracting the optimal point x^* from both sides of the relation in part (a) (see (11)), and by adding and subtracting $\sum_{i=1}^n [\phi_{k+1}]_i \alpha n [\pi_k]_i \nabla f(\hat{x}^k)$, we obtain

$$\hat{x}^{k+1} - x^* = \hat{x}^k - x^* - \sum_{i=1}^n [\phi_{k+1}]_i \alpha n [\pi_k]_i \nabla f(\hat{x}^k) + \alpha \sum_{i=1}^n [\phi_{k+1}]_i \left(n[\pi_k]_i \nabla f(\hat{x}^k) - y_i^k \right).$$

Therefore, by the convexity of the norm and the fact that ϕ_{k+1} is stochastic, we have

$$\|\hat{x}^{k+1}-x^*\| \leq \sum_{i=1}^{n}[\phi_{k+1}]_i\|\hat{x}^k-x^*-\alpha n[\pi_k]_i\nabla f(\hat{x}^k)\|+\alpha\sum_{i=1}^{n}[\phi_{k+1}]_i\|y_i^k-n[\pi_k]_i\nabla f(\hat{x}^k)\|.$$

By Assumption 4 and Assumption 5, the function f is μ-strongly convex and has L-Lipschitz continuous gradients. Thus, for a stepsize α satisfying $\alpha \in (0, \frac{2}{n[\pi_k]_i L})$, for all $i \in [n]$, by Lemma 3.5 it follows that

$$\|\hat{x}^k - x^* - \alpha n[\pi_k]_i\nabla f(\hat{x}^k)\| \leq q_{i,k}(\alpha)\|\hat{x}^k - x^*\|,$$

with $q_{i,k}(\alpha) = \max\{|1-\alpha n[\pi_k]_i\mu|, |1-\alpha n[\pi_k]_i L|\}$.

Let $q_k(\alpha) = \max\{|1-\alpha n\min(\pi_k)\mu|, |1-\alpha n\min(\pi_k)L|\} < 1$, using the preceding relation with the stochasticity of ϕ_{k+1} yields

$$\sum_{i=1}^{n}[\phi_{k+1}]_i\|\hat{x}^k - x^* - \alpha n[\pi_k]_i\nabla f(\hat{x}^k)\| \leq \sum_{i=1}^{n}[\phi_{k+1}]_i q_{i,k}(\alpha)\|\hat{x}^k - x^*\| \leq q_k(\alpha)\|\hat{x}^k - x^*\|.$$

Therefore,

$$\|\hat{x}^{k+1} - x^*\| \leq q_k(\alpha)\|\hat{x}^k - x^*\| + \alpha\sum_{i=1}^{n}[\phi_{k+1}]_i\|y_i^k - n[\pi_k]_i\nabla f(\hat{x}^k)\|. \tag{12}$$

Since $\max(\phi_{k+1}) \leq 1$, to estimate the last term in (12), we factor out $[\pi_k]_i$ as follows

$$\sum_{i=1}^{n}[\phi_{k+1}]_i\|y_i^k-n[\pi_k]_i\nabla f(\hat{x}^k)\| \leq \sum_{i=1}^{n}\|y_i^k-n[\pi_k]_i\nabla f(\hat{x}^k)\| = \sum_{i=1}^{n}[\pi_k]_i\left\|\frac{y_i^k}{[\pi_k]_i}-n\nabla f(\hat{x}^k)\right\|.$$

We add and subtract $\sum_{\ell=1}^{n} y_\ell^k$, and use the triangle inequality for the norm to obtain

$$\sum_{i=1}^{n}[\pi_k]_i\left\|\frac{y_i^k}{[\pi_k]_i} - n\nabla f(\hat{x}^k)\right\| \leq \sum_{i=1}^{n}[\pi_k]_i\left\|\frac{y_i^k}{[\pi_k]_i} - \sum_{\ell=1}^{n}y_\ell^k\right\| + \sum_{i=1}^{n}[\pi_k]_i\left\|\sum_{\ell=1}^{n}y_\ell^k - n\nabla f(\hat{x}^k)\right\|$$

$$\leq \sqrt{\sum_{i=1}^{n}[\pi_k]_i\left\|\frac{y_i^k}{[\pi_k]_i}-\sum_{\ell=1}^{n}y_\ell^k\right\|^2} + \left\|\sum_{\ell=1}^{n}y_\ell^k-n\nabla f(\hat{x}^k)\right\| \leq S(\mathbf{y}^k,\pi_k)+\left\|\sum_{\ell=1}^{n}y_\ell^k-n\nabla f(\hat{x}^k)\right\|.$$

Combining the two preceding relations yields

$$\sum_{i=1}^{n}[\phi_{k+1}]_i\|y_i^k - n[\pi_k]_i\nabla f(\hat{x}^k)\| \leq S(\mathbf{y}^k,\pi_k) + \left\|\sum_{\ell=1}^{n}y_\ell^k - n\nabla f(\hat{x}^k)\right\|. \tag{13}$$

By Lemma 3.2, $\sum_{\ell=1}^{n} y_\ell^k = \sum_{\ell=1}^{n}\nabla f_\ell(x_\ell^k)$, hence, in view of $f = \frac{1}{n}\sum_{\ell=1}^{n} f_\ell$, we have

$$\left\|\sum_{\ell=1}^{n}y_\ell^k - n\nabla f(\hat{x}^k)\right\| = \left\|\sum_{\ell=1}^{n}\Big(\nabla f_\ell(x_\ell^k) - \nabla f_\ell(\hat{x}^k)\Big)\right\| \leq \sum_{\ell=1}^{n}\|\nabla f_\ell(x_\ell^k) - \nabla f_\ell(\hat{x}^k)\|.$$

Since each f_i has L-Lipschitz continuous gradients (by Assumption 4), it follows that

$$\left\|\sum_{\ell=1}^{n} y_\ell^k - n\nabla f(\hat{x}^k)\right\| \leq L\sum_{\ell=1}^{n}\|x_\ell^k - \hat{x}^k\| \leq L\sqrt{\frac{n}{\min(\phi_k)}}D(\mathbf{x}^k,\phi_k). \tag{14}$$

Substituting (13) and (14) in relation (12) gives the desired relation in part (b). ■

The condition $q_k(\alpha) < 1$ of Proposition 4.1(b) holds for example when $\alpha \in (0, \frac{2}{nL})$.

4.2. Weighted Dispersion of Agents' x-variables

In this section, we define and analyze the behavior of a ϕ_k-weighted dispersion of the iterates x_i^k, $i \in [n]$, of the method (5) from their weighted average $\hat{x}^k$, i.e.,

$$D(\mathbf{x}^k,\phi_k) = \sqrt{\sum_{j=1}^{n}[\phi_k]_j\|x_j^k - \hat{x}^k\|^2} \qquad \text{for all } k \geq 0, \tag{15}$$

where the stochastic vectors ϕ_k satisfy $\phi_{k+1}^T A_k = \phi_k^T$ and $\mathbf{x}^k = (x_1^k,\ldots,x_n^k)$.

The dispersion $D(\mathbf{x}^k,\phi_k)$ can be interpreted as the ϕ_k-weighted norm of the difference between $\mathbf{x}^k$ and the vector $\hat{\mathbf{x}}^k = (\hat{x}^k,\ldots,\hat{x}^k)$ consisting of n-copies of $\hat{x}^k$, i.e.,

$$D(\mathbf{x}^k,\phi_k) = \|\mathbf{x}^k - \hat{\mathbf{x}}^k\|_{\phi_k}. \tag{16}$$

Using the definition of x_i^{k+1} in (5a), we can write

$$x_i^{k+1} = z_i^k - \alpha y_i^k, \qquad z_i^k = \sum_{j=1}^{n}[A_k]_{ij}x_j^k, \qquad \text{for all } i \in [n] \text{ and all } k \geq 0. \tag{17}$$

Define $\mathbf{x}^{k+1} = (x_1^{k+1},\ldots,x_n^{k+1})$ and, similarly, define $\mathbf{z}^k = (z_1^k,\ldots,z_n^k)$ and $\mathbf{y}^k = (y_1^k,\ldots,y_n^k)$. Then, we can write the preceding relations compactly as follows

$$\mathbf{x}^{k+1} = \mathbf{z}^k - \alpha\mathbf{y}^k \qquad \text{for all } k \geq 0. \tag{18}$$

We start our analysis by recalling the next lemma:

Lemma 4.2 ([11], Lemma 6)**.** *Let $\mathbb{G} = ([n],\mathcal{E})$ be a strongly connected directed graph, and let A be an row-stochastic matrix that is compatible with the graph and has positive diagonal entries, i.e., $A_{ij} > 0$ when $j = i$ and $(j,i) \in \mathcal{E}$, and $A_{ij} = 0$ otherwise. Also, let ϕ be a stochastic vector and let π be a nonnegative vector such that $\pi^T A = \phi^T$.*

Let $x_1,\ldots,x_n \in \mathbb{R}^p$ be a given collection of vectors, and consider the vectors $z_i = \sum_{j=1}^n A_{ij}x_j$ for all $i \in [n]$. Then, we have

$$\sum_{i=1}^{n}\pi_i\|z_i - u\|^2 \leq \sum_{j=1}^{n}\phi_j\|x_j - u\|^2 - \frac{\min(\pi)\,(\min(A^+))^2}{\max^2(\phi)\mathsf{D}(\mathbb{G})\mathsf{K}(\mathbb{G})}\sum_{j=1}^{n}\phi_j\|x_j - \hat{x}_\phi\|^2 \quad \textit{for all } u \in \mathbb{R}^p,$$

where $\mathsf{D}(\mathbb{G})$ and $\mathsf{K}(\mathbb{G})$ are the diameter and the maximal edge-utility of $\mathbb{G}$, respectively.

The relation for the dispersion $D(\mathbf{x}^k, \phi_k)$ is given in the following proposition.

Proposition 4.3. *Let Assumption 1 and Assumption 2 hold. We have for all $k \geq 0$,*

$$D(\mathbf{x}^{k+1}, \phi_{k+1}) \leq c_k D(\mathbf{x}^k, \phi_k) + \alpha \sqrt{\sum_{i=1}^n [\phi_{k+1}]_i \left\| y_i^k - \sum_{j=1}^n [\phi_{k+1}]_j y_j^k \right\|^2},$$

where the scalar $c_k \in (0,1)$ is given by $c_k = \sqrt{1 - \frac{\min(\phi_{k+1})\, a^2}{\max^2(\phi_k)\, \mathsf{D}(\mathbb{G}_k)\mathsf{K}(\mathbb{G}_k)}}$ and ϕ_k are the stochastic vectors from Lemma 3.3

Proof. We define $\mathbf{v}^k = (\sum_{j=1}^n [\phi_{k+1}]_j y_j^k, \ldots, \sum_{j=1}^n [\phi_{k+1}]_j y_j^k)$, for which we can write the relation for the weighted averages in Proposition 4.1(a) in compact form, as follows

$$\hat{\mathbf{x}}^{k+1} = \hat{\mathbf{x}}^k - \alpha \mathbf{v}^k.$$

Upon subtracting the preceding relation and the compact representation of x-iterate process in (18), we obtain

$$\mathbf{x}^{k+1} - \hat{\mathbf{x}}^{k+1} = \mathbf{z}^k - \hat{\mathbf{x}}^k - \alpha \left(\mathbf{y}^k - \mathbf{v}^k \right).$$

Taking the ϕ_{k+1}-norm on both sides of the preceding relation and using the triangle inequality and the positive scaling property of a norm, we obtain

$$\|\mathbf{x}^{k+1} - \hat{\mathbf{x}}^k\|_{\phi_{k+1}} = \|\mathbf{z}^k - \hat{\mathbf{x}}^k - \alpha \left(\mathbf{y}^k - \mathbf{v}^k \right) \|_{\phi_{k+1}} \leq \|\mathbf{z}^k - \hat{\mathbf{x}}^k\|_{\phi_{k+1}} + \alpha \|\mathbf{y}^k - \mathbf{v}^k\|_{\phi_{k+1}}.$$

The left hand side of the preceding relation corresponds to the dispersion $D(\mathbf{x}^{k+1}, \phi_{k+1})$ (see (16)). The terms on the right hand side we write explicitly in terms of the vector components with $z_i^k = \sum_{j=1}^n [A_k]_{ij} x_j^k$ (see (17)), and obtain

$$D(\mathbf{x}^{k+1}, \phi_{k+1}) \leq \sqrt{\sum_{i=1}^n [\phi_{k+1}]_i \|z_i^k - \hat{x}^k\|^2} + \alpha \sqrt{\sum_{i=1}^n [\phi_{k+1}]_i \left\| y_i^k - \sum_{j=1}^n [\phi_{k+1}]_j y_j^k \right\|^2}. \quad (19)$$

Next, we note that the vectors z_i^k, $i \in [n]$, satisfy Lemma 4.2, with $A = A_k$, and $x_i = x_i^k$ for all $i \in [n]$. Moreover, since we have $\phi_{k+1}^T A_k = \phi_k^T$ by Lemma 3.3, Lemma 4.2 applies with $\pi = \phi_{k+1}$, $\phi = \phi_k$ and $\hat{x}_\phi = \hat{x}^k$, which yields

$$\sum_{i=1}^n [\phi_{k+1}]_i \|z_i^k - \hat{x}^k\|^2 \leq \left(1 - \frac{\min(\phi_{k+1})\, a^2}{\max^2(\phi_k)\, \mathsf{D}(\mathbb{G}_k)\mathsf{K}(\mathbb{G}_k)} \right) \sum_{j=1}^n [\phi_k]_j \|x_j^k - \hat{x}^k\|^2, \quad (20)$$

where we use $\min(A_k^+) \geq a$ (see Assumption 2). Therefore,

$$\sqrt{\sum_{i=1}^n [\phi_{k+1}]_i \|z_i^k - \hat{x}^k\|^2} \leq c_k \sqrt{\sum_{j=1}^n [\phi_k]_j \|x_j^k - \hat{x}^k\|^2}, \quad (21)$$

where $c_k = \sqrt{1 - \frac{\min(\phi_{k+1})\, a^2}{\max^2(\phi_k)\, \mathsf{D}(\mathbb{G}_k)\mathsf{K}(\mathbb{G}_k)}}$.

By recognizing the term on the right hand side of (21) corresponds to $D(\mathbf{x}^k, \phi_k)$, and by combining estimate (21) with (19), we obtain the desired relation. ∎

We conclude this section with a result establishing an estimate for $\|\mathbf{x}^{k+1} - \mathbf{x}^k\|$ and $\left\|\sum_{i=1}^n y_i^k\right\|$, which will be soon used in the analysis of the behavior of the y-iterates.

Lemma 4.4. *Let Assumption 1 and Assumption 2 hold. Then, for all $k \geq 0$, we have*

$$\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\| \leq \left(c_k \sqrt{\frac{1}{\min(\phi_{k+1})}} + \sqrt{\frac{1}{\min(\phi_k)}}\right) D(\mathbf{x}^k, \phi_k) + \alpha \|\mathbf{y}^k\|. \tag{22}$$

Additionally, if Assumption 3 and Assumption 4 hold. Then we have for all $k \geq 0$,

$$\left\|\sum_{i=1}^n y_i^k\right\| \leq L \sqrt{\frac{n}{\min(\phi_k)}} \left(\|\hat{x}^k - x^*\| + D(\mathbf{x}^k, \phi_k)\right).$$

Proof. Adding and subtracting $\hat{\mathbf{x}}^k = (\hat{x}^k, \ldots, \hat{x}^k)$, we have

$$\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\| = \left\|\mathbf{x}^{k+1} - \hat{\mathbf{x}}^k + \hat{\mathbf{x}}^k - \mathbf{x}^k\right\| \leq \left\|\mathbf{z}^k - \hat{\mathbf{x}}^k\right\| + \left\|\mathbf{x}^k - \hat{\mathbf{x}}^k\right\| + \alpha\|\mathbf{y}^k\|,$$

where the last inequality follows from the compact representation of x-iterate process (see (18)) and the triangle inequality. By the relation for norms in (2a), it follows that

$$\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\| \leq \sqrt{\frac{1}{\min(\phi_{k+1})}} \left\|\mathbf{z}^k - \hat{\mathbf{x}}^k\right\|_{\phi_{k+1}} + \sqrt{\frac{1}{\min(\phi_k)}} \left\|\mathbf{x}^k - \hat{\mathbf{x}}^k\right\|_{\phi_k} + \alpha\|\mathbf{y}^k\|.$$

We notice that by the relation in (20), we have $\left\|\mathbf{z}^k - \hat{\mathbf{x}}^k\right\|_{\phi_{k+1}} \leq c_k \left\|\mathbf{x}^k - \hat{\mathbf{x}}^k\right\|_{\phi_k}$. Thus, we obtain the first relation in (22) upon using the definition of $D(\mathbf{x}^k, \phi_k)$ in (15).

Now, we consider $\left\|\sum_{i=1}^n y_i^k\right\|$. By Lemma 3.2, we have

$$\left\|\sum_{i=1}^n y_i^k\right\| = \left\|\sum_{i=1}^n \nabla f_i(x_i^k)\right\| = \left\|\sum_{i=1}^n \left(\nabla f_i(x_i^k) - \nabla f_i(x^*)\right)\right\|,$$

where we use the fact that $\sum_{i=1}^n \nabla f_i(x^*) = 0$, which holds since x^* is the solution to problem (1). Therefore, by using the assumption that each f_i has Lipschitz continuous gradients with a Lipschitz constant $L > 0$, we obtain

$$\left\|\sum_{i=1}^n y_i^k\right\| \leq \sum_{i=1}^n \left\|\nabla f_i(x_i^k) - \nabla f_i(x^*)\right\| \leq L \sum_{i=1}^n \|x_i^k - x^*\| = L\sqrt{n}\|\mathbf{x}^k - \mathbf{x}^*\|.$$

Using the relation for norms in (2a), we further obtain

$$\left\|\sum_{i=1}^n y_i^k\right\| \leq L \sqrt{\frac{n}{\min(\phi_k)}} \|\mathbf{x}^k - \mathbf{x}^*\|_{\phi_k}. \tag{23}$$

Applying the relation (7b) with $u_i = x_i^k$, $\gamma_i = [\phi_k]_i$ for all i, and $u = x^*$ yields

$$\sum_{i=1}^{n} [\phi_k]_i \|x_i^k - x^*\|^2 = \|\hat{x}^k - x^*\|^2 + \sum_{i=1}^{n} [\phi_k]_i \|x_i^k - \hat{x}^k\|^2,$$

where $\hat{x}^k = \sum_{\ell=1}^{n} [\phi_k]_\ell x_\ell^k$. Hence,

$$\left\|\mathbf{x}^k - \mathbf{x}^*\right\|_{\phi_k} = \sqrt{\|\hat{x}^k - x^*\|^2 + \sum_{i=1}^{n} [\phi_k]_i \|x_i^k - \hat{x}^k\|^2} \leq \|\hat{x}^k - x^*\| + \sqrt{\sum_{i=1}^{n} [\phi_k]_i \|x_i^k - \hat{x}^k\|^2},$$

where the inequality in the preceding relation follows from $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, which is valid for any $a, b \geq 0$. Therefore, using the definition of $D(\mathbf{x}^k, \phi_k)$ in (15), we have

$$\left\|\mathbf{x}^k - \mathbf{x}^*\right\|_{\phi_k} \leq \|\hat{x}^k - x^*\| + D(\mathbf{x}^k, \phi_k), \tag{24}$$

from which the second desired relation follows by using (23) and (24). ■

4.3. Weighted Dispersion of Scaled Agents' y-variables

In this section, we analyze the behavior of the directions y_i^k generated by the method in (5). A preliminary result that establishes a basic relation corresponding to a column-stochastic matrix B is given in the following lemma.

Lemma 4.5. *Let $\mathbb{G} = ([n], \mathcal{E})$ be a strongly connected directed graph, and let B be an $n \times n$ column-stochastic matrix that is compatible with the graph and has positive diagonal entries, i.e., $B_{ij} > 0$ when $j = i$ and $(j, i) \in \mathcal{E}$, and $B_{ij} = 0$ otherwise. Also, let ν be a stochastic vector with all entries positive, i.e., $\nu_i > 0$ for all $i \in [n]$, and let the vector π be given by $\pi = B\nu$. Let $y_1, \ldots, y_n \in \mathbb{R}^p$ be a given collection of vectors, and consider the vectors $w_i = \sum_{j=1}^{n} B_{ij} y_j$ for all $i \in [n]$. Then, we have*

$$\sqrt{\sum_{i=1}^{n} \pi_i \left\| \frac{w_i}{\pi_i} - \sum_{\ell=1}^{m} y_\ell \right\|^2} \leq \tau \sqrt{\sum_{i=1}^{n} \nu_i \left\| \frac{y_i}{\nu_i} - \sum_{\ell=1}^{n} y_\ell \right\|^2},$$

where the scalar $\tau \in (0, 1)$ is given by $\tau = \sqrt{1 - \frac{\min^2(\nu)\,(\min(B^+))^2}{\max^2(\nu)\max(\pi)\,\mathsf{D}(\mathbb{G})\mathsf{K}(\mathbb{G})}}$, where $\mathsf{D}(\mathbb{G})$ and $\mathsf{K}(\mathbb{G})$ are the diameter and the maximal edge-utility of the graph $\mathbb{G}$, respectively.

Proof. For any $i \in [n]$, by the definition of w_i, we have

$$\|w_i\|^2 = \left\| \sum_{j=1}^{n} B_{ij} y_j \right\|^2 = \left\| \sum_{j=1}^{n} B_{ij} \nu_j \frac{y_j}{\nu_j} \right\|^2.$$

We further expand the squared norm term by using Lemma 6 with $\gamma_j = B_{ij}\nu_j$ and

$u_j = y_j/\nu_j$ for all $j \in [n]$. Hence, we obtain

$$\|w_i\|^2 = \left(\sum_{\ell=1}^n B_{i\ell}\nu_\ell\right)\sum_{j=1}^n B_{ij}\nu_j \left\|\frac{y_j}{\nu_j}\right\|^2 - \frac{1}{2}\sum_{j=1}^n\sum_{\ell=1}^n B_{ij}\nu_j B_{i\ell}\nu_\ell \left\|\frac{y_j}{\nu_j} - \frac{y_\ell}{\nu_\ell}\right\|^2.$$

Recalling the definition of π, i.e., $\pi = B\nu$, we have $\pi_i = \sum_{\ell=1}^n B_{i\ell}\nu_\ell$, so that we have

$$\|w_i\|^2 = \pi_i \sum_{j=1}^n B_{ij}\nu_j \left\|\frac{y_j}{\nu_j}\right\|^2 - \frac{1}{2}\sum_{j=1}^n\sum_{\ell=1}^n B_{ij}\nu_j B_{i\ell}\nu_\ell \left\|\frac{y_j}{\nu_j} - \frac{y_\ell}{\nu_\ell}\right\|^2.$$

Since the matrix B is nonnegative and compatible with a strongly connected graph $\mathbb{G}$, and since the vector ν has all positive entries, it follows that the vector π also has all entries positive. By dividing with π_i both sides of the preceding relation, and then by summing over all i, we obtain

$$\sum_{i=1}^n \pi_i^{-1}\|w_i\|^2 = \sum_{i=1}^n\sum_{j=1}^n B_{ij}\nu_j \left\|\frac{y_j}{\nu_j}\right\|^2 - \frac{1}{2}\sum_{i=1}^n \pi_i^{-1}\sum_{j=1}^n\sum_{\ell=1}^n B_{ij}\nu_j B_{i\ell}\nu_\ell \left\|\frac{y_j}{\nu_j} - \frac{y_\ell}{\nu_\ell}\right\|^2.$$

For the first term on the right hand side of the preceding inequality, we have that

$$\sum_{i=1}^n\sum_{j=1}^n B_{ij}\nu_j \left\|\frac{y_j}{\nu_j}\right\|^2 = \sum_{i=1}^n\sum_{j=1}^n B_{ij}\nu_j^{-1}\|y_j\|^2 = \sum_{j=1}^n\left(\sum_{i=1}^n B_{ij}\right)\nu_j^{-1}\|y_j\|^2 = \sum_{j=1}^n \nu_j^{-1}\|y_j\|^2,$$

where the last equality follows since the matrix B is column-stochastic. Therefore,

$$\sum_{i=1}^n \pi_i^{-1}\|w_i\|^2 = \sum_{j=1}^n \nu_j^{-1}\|y_j\|^2 - \frac{1}{2}\sum_{i=1}^n \pi_i^{-1}\sum_{j=1}^n\sum_{\ell=1}^n B_{ij}\nu_j B_{i\ell}\nu_\ell \left\|\frac{y_j}{\nu_j} - \frac{y_\ell}{\nu_\ell}\right\|^2. \tag{25}$$

We note that the vector π is stochastic since B is column stochastic and ν is a stochastic vector. Hence,

$$\sum_{i=1}^n \pi_i^{-1}\|w_i\|^2 = \sum_{i=1}^n \pi_i\left\|\frac{w_i}{\pi_i}\right\|^2 = \sum_{i=1}^n \pi_i\left\|\frac{w_i}{\pi_i} - \sum_{\ell=1}^m w_\ell\right\|^2 + \left\|\sum_{\ell=1}^n w_\ell\right\|^2,$$

where the last relation is obtained by using relation (7b) with $u = 0$, $u_i = w_i/\pi_i$, and $\gamma_i = \pi_i$. Using a similar line of arguments, since ν is a stochastic vector, we obtain

$$\sum_{j=1}^n \nu_j^{-1}\|y_j\|^2 = \sum_{j=1}^n \nu_j\left\|\frac{y_j}{\nu_j} - \sum_{\ell=1}^m y_\ell\right\|^2 + \left\|\sum_{\ell=1}^n y_\ell\right\|^2.$$

Since B is column-stochastic, we also have $\sum_{\ell=1}^n w_\ell = \sum_{\ell=1}^n y_\ell$, so that by combining

the preceding two relations with (25), we have that

$$\sum_{i=1}^{n}\pi_i\left\|\frac{w_i}{\pi_i}-\sum_{\ell=1}^{m}y_\ell\right\|^2=\sum_{j=1}^{n}\nu_j\left\|\frac{y_j}{\nu_j}-\sum_{\ell=1}^{m}y_\ell\right\|^2-\frac{1}{2}\sum_{i=1}^{n}\pi_i^{-1}\sum_{j=1}^{n}\sum_{\ell=1}^{n}B_{ij}\nu_jB_{i\ell}\nu_\ell\left\|\frac{y_j}{\nu_j}-\frac{y_\ell}{\nu_\ell}\right\|^2. \quad (26)$$

Next, we estimate the second term on the right hand side of (26). By exchanging the order of the summation so that the summation over i is the last in the order, we obtain

$$\begin{aligned}\sum_{i=1}^{n}\pi_i^{-1}\sum_{j=1}^{n}\sum_{\ell=1}^{n}B_{ij}\nu_jB_{i\ell}\nu_\ell\left\|\frac{y_j}{\nu_j}-\frac{y_\ell}{\nu_\ell}\right\|^2&=\sum_{j=1}^{n}\sum_{\ell=1}^{n}\nu_j\nu_\ell\left\|\frac{y_j}{\nu_j}-\frac{y_\ell}{\nu_\ell}\right\|^2\left(\sum_{i=1}^{n}\pi_i^{-1}B_{ij}B_{i\ell}\right)\\&\geq\sum_{j=1}^{n}\sum_{\ell\in\mathcal{N}_j^{\rm in}}\nu_j\nu_\ell\left\|\frac{y_j}{\nu_j}-\frac{y_\ell}{\nu_\ell}\right\|^2\left(\sum_{i=1}^{n}\pi_i^{-1}B_{ij}B_{i\ell}\right).\end{aligned}$$

The graph $\mathbb{G}$ is strongly connected implying that every node j must have a nonempty in-neighbor set $\mathcal{N}_j^{\rm in}$. Moreover, by assumption we have that $B_{jj}>0$ every $j\in[n]$ and $B_{j\ell}>0$ for all $\ell\in\mathcal{N}_j^{\rm in}$. Therefore, it follows that

$$\sum_{i=1}^{n}\pi_i^{-1}B_{ij}B_{i\ell}\geq\pi_j^{-1}B_{jj}B_{j\ell}\geq\pi_j^{-1}\left(\min_{ij:B_{ij}>0}B_{ij}\right)^2\geq\left(\min_{j\in[n]}\pi_j^{-1}\right)\left(\min_{ij:B_{ij}>0}B_{ij}\right)^2.$$

Using the notation $\min(B^+)=\min_{ij:B_{ij}>0}B_{ij}$, we have

$$\begin{aligned}\sum_{i=1}^{n}\pi_i^{-1}\sum_{j=1}^{n}\sum_{\ell=1}^{n}B_{ij}\nu_jB_{i\ell}\nu_\ell\left\|\frac{y_j}{\nu_j}-\frac{y_\ell}{\nu_\ell}\right\|^2&\geq\frac{(\min(B^+))^2}{\max(\pi)}\sum_{j=1}^{n}\sum_{\ell\in\mathcal{N}_j^{\rm in}}\nu_j\nu_\ell\left\|\frac{y_j}{\nu_j}-\frac{y_\ell}{\nu_\ell}\right\|^2\\&\geq\frac{\min^2(\nu)\,(\min(B^+))^2}{\max(\pi)}\sum_{(\ell,j)\in\mathcal{E}}\left\|\frac{y_j}{\nu_j}-\frac{y_\ell}{\nu_\ell}\right\|^2. \quad (27)\end{aligned}$$

We bound the sum $\sum_{(\ell,j)\in\mathcal{E}}\|y_j-y_\ell\|^2$ from below by employing Lemma 3.1. By assumption the graph $\mathbb{G}=([n],\mathcal{E})$ is strongly connected, by Lemma 3.1 it follows that

$$\sum_{(j,\ell)\in\mathcal{E}}\left\|\frac{y_j}{\nu_j}-\frac{y_\ell}{\nu_\ell}\right\|^2\geq\frac{1}{\mathsf{D}(\mathbb{G})\mathsf{K}(\mathbb{G})}\sum_{j=1}^{n}\sum_{\ell=j+1}^{n}\left\|\frac{y_j}{\nu_j}-\frac{y_\ell}{\nu_\ell}\right\|^2=\frac{1}{\mathsf{D}(\mathbb{G})\mathsf{K}(\mathbb{G})}\,\frac{1}{2}\sum_{j=1}^{n}\sum_{\ell=j}^{n}\left\|\frac{y_j}{\nu_j}-\frac{y_\ell}{\nu_\ell}\right\|^2,$$

where $\mathsf{D}(\mathbb{G})$ is the diameter of the graph $\mathbb{G}$, and $\mathsf{K}(\mathbb{G})$ is the maximal edge-utility in the graph $\mathbb{G}$. The preceding relation and relation (27) yield

$$\begin{aligned}\sum_{i=1}^{n}\pi_i^{-1}\sum_{j=1}^{n}\sum_{\ell=1}^{n}B_{ij}\nu_jB_{i\ell}\nu_\ell\left\|\frac{y_j}{\nu_j}-\frac{y_\ell}{\nu_\ell}\right\|^2&\geq\frac{\min^2(\nu)\;(\min(B^+))^2}{\max(\pi)\,\mathsf{D}(\mathbb{G})\mathsf{K}(\mathbb{G})}\,\frac{1}{2}\sum_{j=1}^{n}\sum_{\ell=1}^{n}\left\|\frac{y_j}{\nu_j}-\frac{y_\ell}{\nu_\ell}\right\|^2\\&\geq\frac{\min^2(\nu)\;(\min(B^+))^2}{\max^2(\nu)\max(\pi)\,\mathsf{D}(\mathbb{G})\mathsf{K}(\mathbb{G})}\,\frac{1}{2}\sum_{j=1}^{n}\sum_{\ell=1}^{n}\nu_j\nu_\ell\left\|\frac{y_j}{\nu_j}-\frac{y_\ell}{\nu_\ell}\right\|^2. \quad (28)\end{aligned}$$

To express the last term, since $\langle \nu, \mathbf{1} \rangle = 1$, we we apply relation (7a) with $\gamma_i = \nu_i$ and $u_i = y_i/\nu_i$ for all $i \in [n]$, and thus obtain

$$\frac{1}{2}\sum_{j=1}^{n}\sum_{\ell=1}^{n}\nu_j\nu_\ell\left\|\frac{y_j}{\nu_j}-\frac{y_\ell}{\nu_\ell}\right\|^2=\sum_{i=1}^{n}\nu_i\left\|\frac{y_i}{\nu_i}-\sum_{\ell=1}^{n}y_\ell\right\|^2.$$

By combining the preceding relation with inequality (28), and by substituting the resulting lower bound back in (26), we obtain

$$\sum_{i=1}^{n}\pi_i\left\|\frac{w_i}{\pi_i}-\sum_{\ell=1}^{m}y_\ell\right\|^2\leq\left(1-\frac{\min^2(\nu)\ (\min(B^+))^2}{\max^2(\nu)\max(\pi)\,\mathsf{D}(\mathbb{G})\mathsf{K}(\mathbb{G})}\right)\sum_{i=1}^{n}\nu_i\left\|\frac{y_i}{\nu_i}-\sum_{\ell=1}^{n}y_\ell\right\|^2.$$

which yields the desired relation after taking the square roots. ■

The third quantity that we use to capture the behavior of the AB/Push-Pull method is the π_k-weighted dispersion of the scaled vectors $y_1^k/[\pi_k]_1, \ldots, y_n^k/[\pi_k]_n$, i.e,

$$S(\mathbf{y}^k, \pi_k) = \sqrt{\sum_{i=1}^{n}[\pi_k]_i\left\|\frac{y_i^k}{[\pi_k]_i}-\sum_{\ell=1}^{n}y_\ell^k\right\|^2}, \tag{29}$$

where π_k is the stochastic vector defined in (9), y_i^k are the directions used in method (5) at time k, and $\mathbf{y}^k = (y_1^k, \ldots, y_n^k)$. We note that $S(\mathbf{y}^k, \pi_k)$ can also be interpreted through the π_k-induced norm in the Cartesian product space $\mathbb{R}^p \times \cdots \times \mathbb{R}^p$. Specifically, using the definition of the iterates y_i^{k+1} in (5b), we express y_i^{k+1} as follows:

$$y_i^{k+1} = w_i^k + \nabla f_i(x_i^{k+1}) - \nabla f_i(x_i^k), \qquad \text{with} \quad w_i^k = \sum_{j=1}^{n}[B_k]_{ij}y_j^k. \tag{30}$$

By defining $\mathbf{w}^k = (w_1^k, \ldots, w_n^k)$ and $\mathbf{g}^k = (\nabla f_1(x_1^k), \ldots, \nabla f_n(x_n^k))$, we have

$$\mathbf{y}^{k+1} = \mathbf{w}^k + \mathbf{g}^{k+1} - \mathbf{g}^k \qquad \text{for all } k \geq 0. \tag{31}$$

Viewing $\mathbf{y}^{k+1}$ as the matrix with columns y_i^{k+1}, and similarly $\mathbf{w}^k$ and $\mathbf{g}^k$, we can write

$$\mathbf{y}^{k+1}\mathrm{diag}^{-1}(\pi_{k+1}) = \mathbf{w}^k\mathrm{diag}^{-1}(\pi_{k+1}) + (\mathbf{g}^{k+1} - \mathbf{g}^k)\mathrm{diag}^{-1}(\pi_{k+1}) \quad \text{for all } k \geq 0, \tag{32}$$

where $\mathrm{diag}(u)$ is the diagonal matrix with the vector u entries on its diagonal. With this alternative view of the method, we have

$$S(\mathbf{y}^k, \pi_k) = \|\mathbf{y}^k\mathrm{diag}^{-1}(\pi_k) - \mathbf{s}^k\|_{\pi_k} \quad \text{with } \mathbf{s}^k = (s^k, \ldots, s^k),\ s^k = \sum_{\ell=1}^{n}y_\ell^k. \tag{33}$$

We provide the recursive relation for $S(\mathbf{y}^k, \pi_k)$ in the following proposition.

Proposition 4.6. *Let Assumptions 1-4 hold, we have for all $k \geq 0$,*

$$\|\mathbf{y}^k\|_{\pi_k^{-1}} = \sqrt{S^2(\mathbf{y}^k, \pi_k) + \left\|\sum_{\ell=1}^n y_\ell^k\right\|^2} \leq S(\mathbf{y}^k, \pi_k) + \left\|\sum_{\ell=1}^n y_\ell^k\right\|,$$

$$S(\mathbf{y}^{k+1}, \pi_{k+1}) \leq \tau_k S(\mathbf{y}^k, \pi_k) + \alpha L r_k \|\mathbf{y}^k\| + L r_k \left(c_k \sqrt{\frac{1}{\min(\phi_{k+1})}} + \sqrt{\frac{1}{\min(\phi_k)}}\right) D(\mathbf{x}^k, \phi_k).$$

Here, the scalars $r_k > 0$ and $\tau_k \in (0,1)$ are given by

$$r_k = \sqrt{n} + \frac{1}{\sqrt{\min(\pi_{k+1})}}, \qquad \tau_k = \sqrt{1 - \frac{\min^2(\pi_k)\, b^2}{\max^2(\pi_k) \max(\pi_{k+1})\, \mathsf{D}(\mathbb{G}_k)\mathsf{K}(\mathbb{G}_k)}},$$

where ϕ_k and π_k are the stochastic vectors associated with the matrices A_k and B_k.

Proof. Firstly, we note that under given assumptions, by Lemma 3.4 we have that the stochastic vectors π_k, $k \geq 0$, defined in (9), all have positive entries. Noting that

$$\|\mathbf{y}^k\|^2_{\pi_k^{-1}} = \sum_{i=1}^n \frac{\|y_i^k\|^2}{[\pi_k]_i} = \sum_{i=1}^n [\pi_k]_i \left\|\frac{y_i^k}{[\pi_k]_i}\right\|^2,$$

and using relation (7b) for the weighted average of vectors, where $\gamma_i = [\pi_k]_i$ and $u_i = y_i^k/[\pi_k]_i$ for all i, and $u = 0$, we obtain

$$\sum_{i=1}^n [\pi_k]_i \left\|\frac{y_i^k}{[\pi_k]_i}\right\|^2 = \sum_{i=1}^n [\pi_k]_i \left\|\frac{y_i^k}{[\pi_k]_i} - \sum_{\ell=1}^n y_\ell^k\right\|^2 + \left\|\sum_{\ell=1}^n y_\ell^k\right\|^2 = S^2(\mathbf{y}^k, \pi_k) + \left\|\sum_{\ell=1}^n y_\ell^k\right\|^2,$$

where the last equality is obtained from the definition of $S(\mathbf{y}^k, \pi_k)$ (see (29)). Hence,

$$\|\mathbf{y}^k\|_{\pi_k^{-1}} = \sqrt{S^2(\mathbf{y}^k, \pi_k) + \left\|\sum_{\ell=1}^n y_\ell^k\right\|^2} \leq S(\mathbf{y}^k, \pi_k) + \left\|\sum_{\ell=1}^n y_\ell^k\right\|, \tag{34}$$

where the inequality is obtained by using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, which is valid for any two scalars $a, b \geq 0$. Thus, we have established the first relation of the proposition.

We next proceed to show the relation for $S(\mathbf{y}^{k+1}, \pi_{k+1})$. By (32), we have

$$\mathbf{y}^{k+1}\mathrm{diag}^{-1}(\pi_{k+1}) = \mathbf{w}^k \mathrm{diag}^{-1}(\pi_{k+1}) + (\mathbf{g}^{k+1} - \mathbf{g}^k)\mathrm{diag}^{-1}(\pi_{k+1}) \qquad \text{for all } k \geq 0.$$

By subtracting the vector $\mathbf{s}^{k+1} = (s^{k+1}, \ldots, s^{k+1})$, where $s^{k+1} = \sum_{\ell=1}^n y_\ell^{k+1}$, from both sides of the preceding relation, we have for all $k \geq 0$,

$$\mathbf{y}^{k+1}\mathrm{diag}^{-1}(\pi_{k+1}) - \mathbf{s}^{k+1} = \mathbf{w}^k \mathrm{diag}^{-1}(\pi_{k+1}) - \mathbf{s}^k + (\mathbf{s}^k - \mathbf{s}^{k+1}) + (\mathbf{g}^{k+1} - \mathbf{g}^k)\mathrm{diag}^{-1}(\pi_{k+1}).$$

By taking π_{k+1}-induced norm on both sides of the preceding equality and by using

relation between $S(\mathbf{y}^{k+1}, \pi_{k+1})$ and the π_{k+1}-induced norm (see (33)), we have that

$$\begin{aligned} S(\mathbf{y}^{k+1}, \pi_{k+1}) &= \|\mathbf{w}^k \mathrm{diag}^{-1}(\pi_{k+1}) - \mathbf{s}^k + (\mathbf{s}^k - \mathbf{s}^{k+1}) + (\mathbf{g}^{k+1} - \mathbf{g}^k)\mathrm{diag}^{-1}(\pi_{k+1})\|_{\pi_{k+1}} \\ &\leq \|\mathbf{w}^k \mathrm{diag}^{-1}(\pi_{k+1}) - \mathbf{s}^k\|_{\pi_{k+1}} + \|\mathbf{s}^k - \mathbf{s}^{k+1}\|_{\pi_{k+1}} + \|(\mathbf{g}^{k+1} - \mathbf{g}^k)\mathrm{diag}^{-1}(\pi_{k+1})\|_{\pi_{k+1}}. \quad (35) \end{aligned}$$

We next consider $\|\mathbf{w}^k \mathrm{diag}^{-1}(\pi_{k+1}) - \mathbf{s}^k\|_{\pi_{k+1}}$, for which by using the definitions of $\mathbf{w}^k$ and $\mathbf{s}^k$, i.e., $\mathbf{w}^k = (w_1^k, \ldots, w_n^k)$ and $\mathbf{s}^k = (s^k, \ldots, s^k)$, we have that

$$\|\mathbf{w}^k \mathrm{diag}^{-1}(\pi_{k+1}) - \mathbf{s}^k\|_{\pi_{k+1}} = \sqrt{\sum_{i=1}^n [\pi_{k+1}]_i \left\| \frac{w_i^k}{[\pi_{k+1}]_i} - s^k \right\|^2},$$

where $s^k = \sum_{\ell=1}^n y_\ell^k$ (see (33)). We now apply Lemma 4.5 with the following identification $\mathbb{G} = \mathbb{G}_k$, $B = B_k$, $\pi = \pi_{k+1}$, and $\nu = \pi_k$, which yields

$$\sum_{i=1}^n [\pi_{k+1}]_i \left\| \frac{w_i^k}{[\pi_{k+1}]_i} - \sum_{\ell=1}^m y_\ell \right\| \leq \tau_k \sqrt{\sum_{i=1}^n [\pi_k]_i \left\| \frac{y_i}{[\pi_k]_i} - \sum_{\ell=1}^n y_\ell \right\|^2},$$

with $\tau_k = \sqrt{1 - \frac{\min^2(\pi_k)\, b^2}{\max^2(\pi_k) \max(\pi_{k+1})\, \mathsf{D}(\mathbb{G}_k)\mathsf{K}(\mathbb{G}_k)}}$, where we use $\min(B_k^+) \geq b$. Hence,

$$\|\mathbf{w}^k \mathrm{diag}^{-1}(\pi_{k+1}) - \mathbf{s}^k\|_{\pi_{k+1}} \leq \tau_k \sqrt{\sum_{i=1}^n [\pi_k]_i \left\| \frac{y_i}{[\pi_k]_i} - \sum_{\ell=1}^n y_\ell \right\|^2} = \tau_k\, S(\mathbf{y}^k, \pi_k),$$

where the equality follows from the definition of $S(\mathbf{y}^k, \pi_k)$ in (29). Thus, by substituting the preceding relation back in (35), we have

$$S(\mathbf{y}^{k+1}, \pi_{k+1}) \leq \tau_k S(\mathbf{y}^k, \pi_k) + \|\mathbf{s}^k - \mathbf{s}^{k+1}\|_{\pi_{k+1}} + \|(\mathbf{g}^{k+1} - \mathbf{g}^k)\mathrm{diag}^{-1}(\pi_{k+1})\|_{\pi_{k+1}}. \quad (36)$$

Next, we consider the term $\|\mathbf{s}^k - \mathbf{s}^{k+1}\|_{\pi_{k+1}}$ in (36), for which we have

$$\|\mathbf{s}^k - \mathbf{s}^{k+1}\|_{\pi_{k+1}} = \sqrt{\sum_{i=1}^n [\pi_{k+1}]_i \|s^{k+1} - s^k\|^2} = \|s^{k+1} - s^k\|,$$

where the last equality follows since the vector π_{k+1} is stochastic. By the definition of s^k in (33), we have $s^k = \sum_{\ell=1}^n y_\ell^k$. Since B_k is column-stochastic, by Lemma 3.2(a) we further have $\sum_{\ell=1}^n y_\ell^k = \sum_{\ell=1}^n \nabla f_\ell(x_\ell^k)$, implying that

$$\|\mathbf{s}^k - \mathbf{s}^{k+1}\|_{\pi_{k+1}} = \left\| \sum_{\ell=1}^n \left(\nabla f_\ell(x_\ell^{k+1}) - \nabla f_\ell(x_\ell^k) \right) \right\| \leq \sum_{\ell=1}^n \|\nabla f_\ell(x_\ell^{k+1}) - \nabla f_\ell(x_\ell^k)\|.$$

By using the Lipschitz continuity of the gradients ∇f_i, we obtain

$$\|\mathbf{s}^k - \mathbf{s}^{k+1}\|_{\pi_{k+1}} \leq L \sum_{\ell=1}^n \|x_\ell^{k+1} - x_\ell^k\| \leq L\sqrt{n} \|\mathbf{x}^{k+1} - \mathbf{x}^k\|. \quad (37)$$

For the term $\|(\mathbf{g}^{k+1}-\mathbf{g}^k)\mathrm{diag}^{-1}(\pi_{k+1})\|_{\pi_{k+1}}$ in relation (36), we have

$$\|(\mathbf{g}^{k+1}-\mathbf{g}^k)\mathrm{diag}^{-1}(\pi_{k+1})\|_{\pi_{k+1}} = \|\mathbf{g}^{k+1}-\mathbf{g}^k\|_{\pi_{k+1}^{-1}} = \sqrt{\sum_{i=1}^{n}\frac{\|\nabla f_i(x_i^{k+1})-\nabla f_i(x_i^k)\|^2}{[\pi_{k+1}]_i}}.$$

By the Lipschitz continuity property of the gradients $\nabla f_i(\cdot)$, we obtain

$$\|(\mathbf{g}^{k+1}-\mathbf{g}^k)\mathrm{diag}^{-1}(\pi_{k+1})\|_{\pi_{k+1}} \leq L\sqrt{\sum_{i=1}^{n}\frac{\|x_i^{k+1}-x_i^k\|^2}{[\pi_{k+1}]_i}} \leq \frac{L}{\sqrt{\min(\pi_{k+1})}}\,\|\mathbf{x}^{k+1}-\mathbf{x}^k\|. \quad (38)$$

Now, we combine the estimates in (37) and (38) with relation (36) and obtain that

$$S(\mathbf{y}^{k+1},\pi_{k+1}) \leq \tau_k\, S(\mathbf{y}^k,\pi_k) + Lr_k\|\mathbf{x}^{k+1}-\mathbf{x}^k\|,$$

where $r_k = \sqrt{n} + \frac{1}{\sqrt{\min(\pi_{k+1})}}$. The desired relation follows from the preceding relation and the estimate for $\|\mathbf{x}^{k+1}-\mathbf{x}^k\|$ in Lemma 4.4 (see (22)). ■

5. Convergence Results

In this section, we combine the results obtained in Sections 4.1–4.3 to obtain a composite relation for the main quantities of interest.

5.1. Composite Relation

We first give the relations in a compact form by defining the vector V_k as follows

$$V_k = \Big(\|\hat{x}^k - x^*\|, D(\mathbf{x}^k,\phi_k), S(\mathbf{y}^k,\pi_k)\Big)^T, \quad (39)$$

which we recall below for convenience:

$$D(\mathbf{x}^k,\phi_k) = \sqrt{\sum_{i=1}^{n}[\phi_k]_i\|x_i^k-\hat{x}^k\|^2}, \quad S(\mathbf{y}^k,\pi_k) = \sqrt{\sum_{j=1}^{n}[\pi_k]_j\left\|\frac{y_j^k}{[\pi_k]_j}-\sum_{\ell=1}^{n}y_\ell^k\right\|^2},$$

where $\hat{x}^k = \sum_{i=1}^{n}[\phi_k]_i x_i^k$ and x^* is the solution of the problem (1). Using Propositions 4.1(c), 4.3, and 4.6, we establish a relation between V_{k+1} and V_k that will involve the constants c_k, and τ_k from Proposition 4.3 and Proposition 4.6, given by

$$q_k(\alpha) = \max\big\{|1-\alpha n\min(\pi_k)\mu|, |1-\alpha n\min(\pi_k)L|\big\},\ r_k = \sqrt{n}+\frac{1}{\sqrt{\min(\pi_{k+1})}}, \quad (40a)$$

$$c_k = \sqrt{1-\frac{\min(\phi_{k+1})\,a^2}{\max^2(\phi_k)\mathsf{D}(\mathbb{G}_k)\mathsf{K}(\mathbb{G}_k)}}, \tau_k = \sqrt{1-\frac{\min^2(\pi_k)\,b^2}{\max^2(\pi_k)\max(\pi_{k+1})\mathsf{D}(\mathbb{G}_k)\mathsf{K}(\mathbb{G}_k)}}. \quad (40b)$$

For the vector V_k we have the following result.

Proposition 5.1. *Let Assumptions 1-5 hold. Consider the iterates produced by the AB/Push-Pull method in (5) with the stepsize $\alpha \in (0, 2(nL)^{-1})$, we have*

$$V_{k+1} \leq M_k(\alpha) V_k \qquad \text{for all } k \geq 0,$$

where $M_k(\alpha)$ is the matrix given by

$$M_k(\alpha) = \begin{bmatrix} q_k(\alpha) & \alpha L\sqrt{n}\varphi_k & \alpha \\ \alpha L\gamma_k\sqrt{n}\varphi_k & c_k + \alpha L\gamma_k\sqrt{n}\varphi_k & \alpha\gamma_k \\ \alpha L^2 r_k\sqrt{n}\varphi_k & Lr_k(c_k\varphi_{k+1} + \varphi_k) + \alpha L^2 r_k\sqrt{n}\varphi_k & \tau_k + \alpha L r_k \end{bmatrix}.$$

with $\gamma_k = \sqrt{\max_{j\in[n]}([\phi_{k+1}]_j[\pi_k]_j)}$, $\varphi_k = \sqrt{\frac{1}{\min(\phi_k)}}$, and $q_k(\alpha)$, c_k, τ_k, and r_k as in (40).

Proof. The first row of $M_k(\alpha)$ is given by Proposition 4.1(b) when $\alpha \in (0, 2(nL)^{-1})$. Next, we consider the relation for $D(\mathbf{x}^{k+1}, \phi_{k+1})$. By Proposition 4.3, we have that

$$D(\mathbf{x}^{k+1}, \phi_{k+1}) \leq c_k D(\mathbf{x}^k, \phi_k) + \alpha \sqrt{\sum_{i=1}^n [\phi_{k+1}]_i \left\| y_i^k - \sum_{j=1}^n [\phi_{k+1}]_j y_j^k \right\|^2}. \tag{41}$$

Using relation (7b) with $\gamma_i = [\phi_{k+1}]_i$, $u_i = y_i^k$ for all i and $u = 0$, it follows that

$$\sum_{i=1}^n [\phi_{k+1}]_i \left\| y_i^k - \sum_{j=1}^n [\phi_{k+1}]_j y_j^k \right\|^2 \leq \sum_{i=1}^n [\phi_{k+1}]_i \left\| y_i^k \right\|^2.$$

By multiplying and dividing each term in the summation on the right hand side with $[\pi_k]_i$, we find that

$$\sum_{i=1}^n [\phi_{k+1}]_i \left\| y_i^k - \sum_{j=1}^n [\phi_{k+1}]_j y_j^k \right\|^2 \leq \sum_{i=1}^n [\phi_{k+1}]_i [\pi_k]_i \frac{\left\| y_i^k \right\|^2}{[\pi_k]_i} \leq \max_{j\in[n]}([\phi_{k+1}]_j[\pi_k]_j) \sum_{i=1}^n \frac{\left\| y_i^k \right\|^2}{[\pi_k]_i}.$$

Therefore, by taking the square roots on both sides of the preceding relation, we obtain

$$\sqrt{\sum_{i=1}^n [\phi_{k+1}]_i \left\| y_i^k - \sum_{j=1}^n [\phi_{k+1}]_j y_j^k \right\|^2} \leq \sqrt{\max_{j\in[n]}([\phi_{k+1}]_j[\pi_k]_j)} \sqrt{\sum_{i=1}^n \frac{\left\| y_i^k \right\|^2}{[\pi_k]_i}} = \gamma_k \|\mathbf{y}^k\|_{\pi_k^{-1}},$$

where the equality follows upon using $\gamma_k = \sqrt{\max_{j\in[n]}([\phi_{k+1}]_j[\pi_k]_j)}$ and the definition of $\|\mathbf{y}^k\|_{\pi_k^{-1}}$. Substituting the preceding estimate back in relation (41), we find that,

$$D(\mathbf{x}^{k+1}, \phi_{k+1}) \leq c_k D(\mathbf{x}^k, \phi_k) + \alpha\gamma_k \|\mathbf{y}^k\|_{\pi_k^{-1}} \qquad \text{for all} \qquad k \geq 0.$$

Using the preceding relation, the relation $\|\mathbf{y}^k\|_{\pi_k^{-1}} \leq S(\mathbf{y}^k, \pi_k) + \left\|\sum_{\ell=1}^n y_\ell^k\right\|$ established in Proposition 4.6, and the following relation from Lemma 4.4

$$\left\|\sum_{i=1}^n y_i^k\right\| \leq L\sqrt{\frac{n}{\min(\phi_k)}}\left(\|\hat{x}^k - x^*\| + D(\mathbf{x}^k, \phi_k)\right), \tag{42}$$

we obtain the desired relation for $D(\mathbf{x}^{k+1}, \phi_{k+1})$ (given by the second row of $M_k(\alpha)$).

Lastly, the relation for $S(\mathbf{y}^{k+1}, \pi_{k+1})$ comes from Proposition 4.6. For the quantity $\|\mathbf{y}^k\|$, using the vector-induced norm property in (2b) and the fact that the vector π_k is stochastic for all k, we have $\|\mathbf{y}^k\| \leq \|\mathbf{y}^k\|_{\pi_k^{-1}}$. Upon using the relations $\|\mathbf{y}^k\|_{\pi_k^{-1}} \leq S(\mathbf{y}^k, \pi_k) + \left\|\sum_{\ell=1}^n y_\ell^k\right\|$ established in Proposition 4.6 and (42), we obtain

$$\begin{aligned} S(\mathbf{y}^{k+1}, \pi_{k+1}) \leq{}& (\tau_k + \alpha L r_k)\ S(\mathbf{y}^k, \pi_k) + \alpha L^2 r_k \sqrt{\frac{n}{\min(\phi_k)}}\|\hat{x}^k - x^*\| \\ &+ L r_k \left(c_k\sqrt{\frac{1}{\min(\phi_{k+1})}} + \sqrt{\frac{1}{\min(\phi_k)}} + \alpha L\sqrt{\frac{n}{\min(\phi_k)}}\right) D(\mathbf{x}^k, \phi_k), \end{aligned}$$

which gives the third row of $M_k(\alpha)$. ■

5.2. Convergence Result and Range for the Stepsize

From Proposition 5.1, to prove that V_k tends to 0 at a geometric rate, it is sufficient to show that $M_k(\alpha) \leq M(\alpha)$ for some matrix $M(\alpha)$, and then choose a suitable stepsize $\alpha \in (0, 2(nL)^{-1})$ such that the eigenvalues of $M(\alpha)$ are inside the unit circle, i.e., the spectral radius of $M(\alpha)$ is less than 1.

We now determine an upper bound matrix $M(\alpha)$ for $M_k(\alpha)$. Let $c \in (0,1)$, $\tau \in (0,1)$, r, and φ, be upper bounds for c_k, τ_k, r_k, and φ_k, respectively, i.e.,

$$\max_{k\geq 0} c_k \leq c, \quad \max_{k\geq 0} \tau_k \leq \tau, \quad \max_{k\geq 0} r_k \leq r, \quad \max_{k\geq 0} \varphi_k \leq \varphi. \tag{43}$$

For the quantity $q_k(\alpha)$ as in (40a), when $\alpha \in (0, 2(nL + n\mu)^{-1})$, we have $q_k(\alpha) = 1 - \alpha n \min(\pi_k)\mu < 1$. Let σ be a lower bound for $\min(\pi_k)$, $k \geq 0$, corresponding to the graph sequence $\{\mathbb{G}_k\}$. In the most general case of graph sequences, by Lemma 3.4 we have that $\sigma \leq \min_{k\geq 0}\{\min(\pi_k)\}$ with $\sigma \geq \frac{b^n}{n} > 0$. Thus, we have the following upper bound for $q_k(\alpha)$:

$$\max_{k\geq 0} q_k(\alpha) \leq 1 - \alpha n \sigma \mu \in (0,1) \qquad \text{where} \qquad \sigma \leq \min_{k\geq 0}\{\min(\pi_k)\}. \tag{44}$$

We notice also that $\gamma_k = \max_{j\in[n]}([\phi_{k+1}]_j[\pi_k]_j) \leq 1$ since ϕ_k and π_k are stochastic vectors. Using these upper-bounds, for $\alpha \in (0, 2(nL+n\mu)^{-1})$, we have $M_k(\alpha) \leq M(\alpha)$, for all $k \geq 0$, with the matrix $M(\alpha)$ given by

$$M(\alpha) = \begin{bmatrix} 1 - \alpha n\sigma\mu & \alpha L\sqrt{n}\varphi & \alpha \\ \alpha L\sqrt{n}\varphi & c + \alpha L\sqrt{n}\varphi & \alpha \\ \alpha L^2 r\sqrt{n}\varphi & Lr(1+c)\varphi + \alpha L^2 r\sqrt{n}\varphi & \tau + \alpha L r \end{bmatrix}. \tag{45}$$

Proposition 5.2. *Let Assumptions 1-5 hold. Consider the iterates produced by the method in (5) and the notation in (43)-(44). If the stepsize $\alpha > 0$ is chosen such that*

$$\alpha \leq \min\left\{\frac{1-c}{L\sqrt{n}\varphi},\ \frac{1-\tau}{Lr},\ \frac{n\sigma\mu(1-\tau)(1-c)}{\eta},\frac{2}{n(L+\mu)}\right\}, \tag{46}$$

where $\eta = L(n\sigma\mu + L\sqrt{n}\varphi)\left((1+c)r\varphi + (1-c)r + (1-\tau)\sqrt{n}\varphi\right) > 0$. Then,

$$\lim_{k\to\infty} \|x_i^k - x^*\| = 0, \qquad \textit{for all} \quad i \in [n].$$

Proof. Recall that by Proposition 5.1, we have $V_{k+1} \leq M_k(\alpha)V_k$, for all $k \geq 0$. Therefore, with the matrix $M(\alpha)$ defined as in (45), we have

$$V_{k+1} \leq M(\alpha)V_k, \qquad \text{for all } k \geq 0. \tag{47}$$

Thus, $\|\hat{x}^k - x^*\|$, $D(\mathbf{x}^k, \phi_k)$ and $S(\mathbf{y}^k, \pi_k)$ all converge to 0 linearly at rate $\mathcal{O}\left(\rho_M^k\right)$ if the spectral radius $\rho_{M(\alpha)}$ of $M(\alpha)$ satisfies $\rho_{M(\alpha)} < 1$. By Lemma 8 of [17], we will have $\rho_{M(\alpha)} < 1$ if all diagonal entries of $M(\alpha)$ are less than 1 and $\det(\mathbb{I} - M(\alpha)) > 0$, where

$$\det(M(\alpha) - \mathbb{I}) = \begin{vmatrix} -\alpha n\sigma\mu & \alpha L\sqrt{n}\varphi & \alpha \\ \alpha L\sqrt{n}\varphi & c + \alpha L\sqrt{n}\varphi - 1 & \alpha \\ \alpha L^2 r\sqrt{n}\varphi & Lr(1+c)\varphi + \alpha L^2 r\sqrt{n}\varphi & \tau + \alpha Lr - 1 \end{vmatrix}.$$

Hence,

$$\det(M(\alpha) - \mathbb{I}) = \alpha\left[\alpha\eta - n\sigma\mu(1-\tau)(1-c)\right],$$

where $\eta = L(n\sigma\mu + L\sqrt{n}\varphi)\left((1+c)r\varphi + (1-c)r + (1-\tau)\sqrt{n}\varphi\right) > 0$ since $c < 1$ and $\tau < 1$. It remains to choose $\alpha \in (0, 2(nL + n\mu)^{-1})$ so that the following conditions are satisfied

$$\begin{cases} c + \alpha L\sqrt{n}\varphi < 1 \\ \tau + \alpha Lr < 1 \\ \alpha\eta - n\sigma\mu(1-\tau)(1-c) < 0. \end{cases}$$

Solving the preceding system of inequalities yields the range in (46). ∎

Remark 1. We can relax Assumption 1 by considering a C-strongly-connected graph sequence, i.e., there exists some integer, $C \geq 1$ such that the graph with edge set $\mathcal{E}_k^C = \bigcup_{i=kC}^{(k+1)C-1} \mathcal{E}_i$ is strongly connected for every $k \geq 0$. In this case, the more general results of Lemma 3.3 and Lemma 3.4 state that there exist stochastic vector sequences $\{\phi_k\}$ and $\{\pi_k\}$, such that for all $k \geq 0$,

$$\phi'_{k+C}\left(A_{k+C-1}\dots A_{k+1}A_k\right) = \phi'_k \quad \text{and} \quad \pi_{k+C} = \left(B_{k+C-1}\dots B_{k+1}B_k\right)\pi_k.$$

Furthermore,

$$[\phi_k]_i \geq \frac{a^{nC}}{n} \quad \text{and} \quad [\pi_k]_i \geq \frac{b^{nC}}{n} \qquad \text{for all} \quad i \in [n].$$

With the use of these results, the rest of convergence analysis follows similarly to our analysis for strongly connected graphs, by noticing that contractions due to row- and column-stochastic matrices occur after time $k = C$.

6. Numerical Simulations

In this section, we evaluate the performance of the proposed algorithm through a sensor fusion problem over a network, as described in [37]. The estimation problem is given as follows

$$\min_{x \in \mathbb{R}^p} \sum_{i=1}^{n} \left(\|z_i - H_i x\|^2 + \lambda_i \|x\|^2 \right),$$

where x is the unknown parameter to be estimated, $H_i \in \mathbb{R}^{s \times p}$ represents the measurement matrix, $z_i = H_i x + \omega_i \in \mathbb{R}^s$ is the noisy observation of sensor i with some noise ω_i and λ_i is the regularization parameter for the local cost function of sensor i.

As in [17], we set $n = 20$, $p = 20$ and $s = 1$ so that each local cost function is ill-conditioned, requiring the coordination among agents to achive fast convergence. The measurement matrix H_i is generated from a uniform distribution in the unit $\mathbb{R}^{s \times p}$ space which is then normalized such that its Lipschitz constant is equal to 1. The noise ω_i follows an i.i.d. Gaussian process with zero mean and unit variance $\mathcal{N}(0, 1)$. The regularization parameter is chosen to be $\lambda_i = 0.01$, for all $i \in [n]$, to ensure the strong convexity of the loss function.

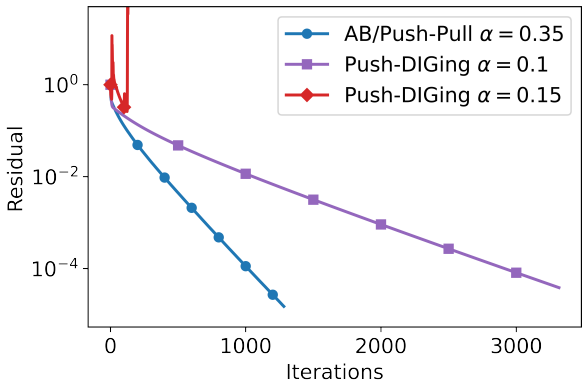

Figure 1: Residuals plot

We compare our proposed AB/Push-Pull algorithm against Push-DIGing [10]. The simulation is carried out over a random sequence of time-varying directed communication network . The performance is compared in terms of the relative residual defined as $\frac{\|\mathbf{x}^k - \mathbf{x}^*\|_2^2}{\|\mathbf{x}^0 - \mathbf{x}^*\|_2^2}$. Figure 1 illustrates the performance of the above algorithms under a randomly generated time-varying network. As discussed in [17], AB/Push-Pull allows for much larger value of the stepsize compared to Push-DIGing and it converges faster especially for ill-conditioned problems and when graphs are not well balanced.

7. Conclusions

In this paper, we study a distributed optimization problem over a time-varying directed communication network. We consider the AB/Push-Pull gradient-based method where each node maintains estimates of the optimal decision variable and the average gradient of the agents' objective functions. The information about the decision variable is pushed to its neighbors, while the information about the gradients is pulled from its neighbors using both row- and column-stochastic weights simultaneously. We explore the contractive properties of the iterates produced by the method, which are inherited from the use of the mixing terms and the fact that the mixing matrices are compliant with a directed strongly connected graph. We prove that the algorithm converges linearly to the global minimizer for smooth and strongly convex cost functions. The convergence result is derived based on the choice of appropriate stepsize values for which explicit upper bounds are provided in terms of the properties of the cost functions, the mixing matrices, and the graph connectivity structure.

References

[1] J. Bazerque and G. Giannakis, *Distributed Spectrum Sensing for Cognitive Radio Networks by Exploiting Sparsity*, IEEE Trans. Signal Process. 58 (2010), pp. 1847–1862.

[2] J. Duchi, A. Agarwal, and M. Wainwright, *Dual Averaging for Distributed Optimization: Convergence Analysis and Network Scaling*, IEEE Trans. Autom. Control 57 (2012), pp. 592–606.

[3] H. Gao, Y. Wang, and A. Nedić, *Dynamics based Privacy Preservation in Decentralized Optimization* (2022). Available at https://arxiv.org/abs/2207.05350.

[4] B. Gharesifard and J. Cortes, *Distributed Strategies for Generating Weight-Balanced and Doubly Stochastic Digraphs*, European Journal of Control 18 (2012), pp. 539–557.

[5] H. Li and Z. Lin, *Accelerated Gradient Tracking over Time-Varying Graphs for Decentralized Optimization* (2021). Available at https://arxiv.org/abs/2104.02596.

[6] A. Mokhtari, Q. Ling, and A. Ribeiro, *Network Newton Distributed Optimization Methods*, IEEE Trans. Signal Process. 65 (2017), pp. 146–161.

[7] A. Nedić and A. Ozdaglar, *Distributed Subgradient Methods for Multi-agent Optimization*, IEEE Trans. Autom. Control 54 (2009), pp. 48–61.

[8] A. Nedić, *Asynchronous Broadcast-Based Convex Optimization over a Network*, IEEE Trans. Autom. Control 56 (2011), pp. 1337–1351.

[9] A. Nedić and A. Olshevsky, *Distributed Optimization over Time-Varying Directed Graphs*, IEEE Trans. Autom. Control 60 (2015), pp. 601–615.

[10] A. Nedić, A. Olshevsky, and W. Shi, *Achieving Geometric Convergence for Distributed Optimization Over Time-Varying Graphs*, SIAM J. Optim. 27 (2017), pp. 2597–2633.

[11] D.T.A. Nguyen, D.T. Nguyen, and A. Nedić, *Distributed Nash Equilibrium Seeking over Time-Varying Directed Communication Networks* (2022). Available at https://arxiv.org/pdf/2201.02323.pdf.

[12] A. Olshevsky, *Linear Time Average Consensus and Distributed Optimization on Fixed Graphs*, SIAM J. Control Optim. 55 (2017), pp. 3990–4014.

[13] B. Polyak, *Introduction to Optimization*, New York : Optimization Software, Inc., 1987.

[14] S. Pu, *A Robust Gradient Tracking Method for Distributed Optimization over Directed Networks*, in *2020 59th IEEE Conf. Decis. Control.* 2020, pp. 2335–2341.

[15] S. Pu and A. Garcia, *A Flocking-based Approach for Distributed Stochastic Optimization*, Operations Research 1 (2018), pp. 267–281.

[16] S. Pu and A. Nedić, *A Distributed Stochastic Gradient Tracking Method*, in *2018 IEEE Conf. Decis. Control (CDC).* 2018, pp. 963–968.

[17] S. Pu, W. Shi, J. Xu, and A. Nedić, *Push–Pull Gradient Methods for Distributed Optimization in Networks*, IEEE Trans. Autom. Control 66 (2021), pp. 1–16.
[18] G. Qu and N. Li, *Harnessing Smoothness to Accelerate Distributed Optimization*, IEEE Trans. Control. Netw. Syst. 5, pp. 1245–1260.
[19] M. Rabbat and R. Nowak, *Distributed Optimization in Sensor Networks*, in *Third International Symposium on Information Processing in Sensor Networks*. 2004, pp. 20–27.
[20] R. Raffard, C. Tomlin, and S. Boyd, *Distributed Optimization for Cooperative Agents: Application to Formation Flight*, in *43rd IEEE Conf. Decis. Control (CDC)*, Vol. 3. 2004, pp. 2453–2459.
[21] S. Ram, V. Veeravalli, and A. Nedić, *Distributed Non-Autonomous Power Control through Distributed Convex Optimization*, in *INFOCOM*. 2009, pp. 3001–3005.
[22] F. Saadatniaki, R. Xin, and U.A. Khan, *Decentralized Optimization Over Time-Varying Directed Graphs With Row and Column-Stochastic Matrices*, IEEE Trans. Autom. Control 65 (2020), pp. 4769–4780.
[23] W. Shi, Q. Ling, G. Wu, and W. Yin, *A Proximal Gradient Algorithm for Decentralized Composite Optimization*, IEEE Trans. Signal Process. 63 (2015), pp. 6013–6023.
[24] W. Shi, Q. Ling, G. Wu, and W. Yin, *EXTRA: An Exact First-Order Algorithm for Decentralized Consensus Optimization*, SIAM J. Optim. 25 (2015), pp. 944–966.
[25] W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin, *On the Linear Convergence of the ADMM in Decentralized Consensus Optimization*, IEEE Trans. Signal Process. 62 (2014), pp. 1750–1761.
[26] K. Srivastava and A. Nedić, *Distributed Asynchronous Constrained Stochastic Optimization*, IEEE Journal of Selected Topics in Signal Processing 5 (2011), pp. 772–790.
[27] D. Stipanovic, G. Inalhan, R. Teo, and C. Tomlin, *Decentralized Overlapping Control of A Formation of Unmanned Aerial Vehicles*, in *41st IEEE Conf. Decis. Control*, Vol. 3. 2002, pp. 2829–2835 vol.3.
[28] K. Tsianos, S. Lawlor, and M. Rabbat, *Push-Sum Distributed Dual Averaging for Convex Optimization*, in *Proceedings of the 51st IEEE Conf. Decis. Control*. 2012, pp. 5453–5458.
[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*. 2012, pp. 1543–1550.
[30] D. Varagnolo, F. Zanella, A. Cenedese, G. Pillonetto, and L. Schenato, *Newton-Raphson Consensus for Distributed Convex Optimization*, IEEE Trans. Autom. Control 61 (2016), pp. 994–1009.
[31] Y. Wang and A. Nedić, *Tailoring Gradient Methods for Differentially-Private Distributed Optimization* (2022). Available at https://arxiv.org/abs/2202.01113.
[32] C. Xi, V.S. Mai, R. Xin, E.H. Abed, and U.A. Khan, *Linear Convergence in Optimization over Directed Graphs with Row-Stochastic Matrices*, IEEE Trans. Autom. Control (2018).
[33] C. Xi, R. Xin, and U.A. Khan, *ADD-OPT: Accelerated Distributed Directed Optimization*, IEEE Trans. Autom. Control 63 (2018), pp. 1329–1339.
[34] R. Xin, A.K. Sahu, U.A. Khan, and S. Kar, *Distributed Stochastic Optimization with Gradient Tracking over Strongly-connected Networks*, in *58th IEEE Conf. Decis. Control, Nice, France*. 2019, pp. 8353–8358.
[35] R. Xin and U.A. Khan, *A Linear Algorithm for Optimization Over Directed Graphs With Geometric Convergence*, IEEE Contr. Syst. Lett. 2 (2018), pp. 315–320.
[36] R. Xin, C. Xi, and U.A. Khan, *FROST—Fast Row-Stochastic Optimization with Uncoordinated Step-sizes*, EURASIP J. Adv. Signal Process. (2019), pp. 1–14.
[37] J. Xu, S. Zhu, Y.C. Soh, and L. Xie, *Convergence of Asynchronous Distributed Gradient Methods Over Stochastic Networks*, IEEE Trans. Autom. Control 63 (2018), pp. 434–448.