Distributed stochastic optimization with gradient tracking over strongly-connected networks

Ran Xin, Anit Kumar Sahu, Usman A. Khan and Soummya Kar

Abstract

In this paper, we study distributed stochastic optimization to minimize a sum of smooth and strongly-convex local cost functions over a network of agents, communicating over a strongly-connected graph. Assuming that each agent has access to a stochastic first-order oracle ($\mathcal{SFO}$), we propose a novel distributed method, called $\mathcal{S}$-$\mathcal{AB}$, where each agent uses an auxiliary variable to asymptotically track the gradient of the global cost in expectation. The $\mathcal{S}$-$\mathcal{AB}$ algorithm employs row- and column-stochastic weights simultaneously to ensure both consensus and optimality. Since doubly-stochastic weights are not used, $\mathcal{S}$-$\mathcal{AB}$ is applicable to arbitrary strongly-connected graphs. We show that under a sufficiently small constant step-size, $\mathcal{S}$-$\mathcal{AB}$ converges linearly (in expected mean-square sense) to a neighborhood of the global minimizer. We present numerical simulations based on real-world data sets to illustrate the theoretical results.

Index Terms

Stochastic optimization, first-order methods, multi-agent systems, directed graphs

I. INTRODUCTION

In the era of data deluge, where it is particularly difficult to store and process all data on a single device/node/processor, distributed schemes are becoming attractive for inference, learning, and optimization. Distributed optimization over multi-agent systems, thus, has been of significant interest in many areas including but not limited to machine learning [1], [2], big-data analytics [3], [4], and distributed control [5], [6]. However, the underlying algorithms must be designed to address practical limitations and realistic scenarios. For instance, with the computation and data collection/storage being pushed to the edge devices, e.g., in Internet of Things (IoT), the data available for distributed optimization is often inexact. Moreover, the ad hoc nature of setups outside of data centers requires the algorithms to be amenable to communication protocols that are not necessarily bidirectional. The focus of this paper is to study and characterize distributed optimization schemes where the inter-agent communication is restricted to directed graphs and the information/data is inexact.

In particular, we study distributed stochastic optimization over directed graphs and propose the $\mathcal{S}$-$\mathcal{AB}$ algorithm to minimize a sum of local cost functions. The $\mathcal{S}$-$\mathcal{AB}$ algorithm assumes access to a stochastic first-order oracle ($\mathcal{SFO}$), i.e., when an agent queries the $\mathcal{SFO}$, it gets an unbiased estimate of the gradient of its local cost function. In the proposed approach, each agent makes a weighted average of its own and its neighbors solution estimates, and simultaneously incorporates its local gradient estimate of the global cost function. The exchange of solution estimates is performed over a row-stochastic weight matrix. In parallel, each agent maintains its own estimate of the gradient of the global cost function, by simultaneously incorporating a weighted average of its and its neighbors' gradient estimates and its local gradient tracking estimate. The exchange of gradient estimates of the global cost function is performed over a column-stochastic weight matrix. Since doubly-stoachstic weights are nowhere used, $\mathcal{S}$-$\mathcal{AB}$ is an attractive solution that is applicable to arbitrary, strongly-connected graphs.

The main contributions of this paper are as follows: (i) We show that, by choosing a sufficiently small constant step-size, α, $\mathcal{S}$-$\mathcal{AB}$ converges linearly to a neighborhood of the global minimizer. This convergence guarantee is achieved for continuously-differentiable, strongly-convex, local cost functions, where each agent is assumed to have access to a $\mathcal{SFO}$ and the gradient noise has zero-mean and bounded variance. (ii) We provide explicit expressions of the appropriate norms under which the row- and column-stochastic weight matrices contract. With the help of these norms, we develop sharp and explicit convergence arguments.

We now briefly review the literature concerning distributed and stochastic optimization. Early work on deterministic finite-sum problems includes [7]–[9], while work on stochastic problems can be found in [10], [11]. Recently, gradient tracking has been proposed where the local gradient at each agent is replaced by the estimate of the global gradient [12]–[15]. Methods for directed graphs that are based on gradient tracking [14]–[21] rely on separate iterations for eigenvector estimation that may impede the convergence. This issue was recently resolved in [22], [23], see also [24]–[27] for the follow-up work, where eigenvector estimation was removed with the help of a unique approach that uses both row- and column-stochastic weights. Ref. [22] derives linear convergence of the finite-sum problem when the underlying functions are smooth and strongly-convex, however, since arbitrary norms are used in the analysis, the convergence bounds are not sharp. Recent related work on time-varying networks and other approaches can be found in [28]–[32], albeit, without gradient tracking. Of significant relevance is [33], where a similar setup with gradient tracking is considered over undirected graphs. We note that $\mathcal{S}$-$\mathcal{AB}$ generalizes [33] and the analysis in [33] relies on the weight matrix contraction in 2-norm that is not applicable here.

We now describe the rest of the paper. Section II describes the problem, assumptions, and some auxiliary results. We present the convergence analysis in Section III and the main result in Section IV. Finally, Section V provides the numerical experiments and Section VI concludes the paper.

Basic Notation: We use lowercase bold letters for vectors and uppercase italic letters for matrices. We use I_n for the $n \times n$ identity matrix, and $\mathbf{1}_n$ for the column of n ones. For an arbitrary vector, $\mathbf{x}$, we denote its ith element by $[\mathbf{x}]_i$ and its smallest element by $\underline{\mathbf{x}}$ and its largest element by $\overline{\mathbf{x}}$. Inequalities involving matrices and

vectors are to be interpreted componentwise. For a matrix, X, we denote $\rho(X)$ as its spectral radius and X_∞ as its infinite power (if it exists), i.e., $X_\infty = \lim_{k \to \infty} X^k$. For a primitive, row-stochastic matrix, A, we denote its left and right eigenvectors corresponding to the eigenvalue of 1 by $\boldsymbol{\pi}_r$ and $\mathbf{1}_n$, respectively, such that $\boldsymbol{\pi}_r^\top \mathbf{1}_n = 1$ and $A_\infty = \mathbf{1}_n \boldsymbol{\pi}_r^\top$. Similarly, for a primitive, column-stochastic matrix, B, we have $B_\infty = \boldsymbol{\pi}_c \mathbf{1}_n^\top$.

II. PROBLEM FORMULATION AND AUXILIARY RESULTS

Consider n agents connected over a directed graph, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, \cdots, n\}$ 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. We assume that $(i,i) \in \mathcal{E}, \forall i$. The agents solve the following problem:

$$\text{P1}: \quad \min_{\mathbf{x} \in \mathbb{R}^p} F(\mathbf{x}) \triangleq \frac{1}{n} \sum_{i=1}^n f_i(\mathbf{x}), \tag{1}$$

where each $f_i : \mathbb{R}^p \to \mathbb{R}$ is known only to agent i. We now formalize the assumptions.

Assumption 1: Each local objective, f_i, is μ-strongly-convex, i.e., $\forall i \in \mathcal{V}$ and $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^p$. Thus, we have

$$f_i(\mathbf{y}) \geq f_i(\mathbf{x}) + \nabla f_i(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{\mu}{2} \|\mathbf{x} - \mathbf{y}\|_2^2.$$

Under Assumption 1, the optimal solution for Problem P1 exists and is unique, which we denote as $\mathbf{x}^*$.

Assumption 2: Each local objective, f_i, is l-smooth, i.e., its gradient is Lipschitz-continuous: $\forall i \in \mathcal{V}$ and $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^p$, we have, for some $l > 0$,

$$\|\nabla f_i(\mathbf{x}) - \nabla f_i(\mathbf{y})\|_2 \leq l \|\mathbf{x} - \mathbf{y}\|_2.$$

We make the following assumption on the agent communication graph, which guarantees the existence of a *directed* path from each agent i to each agent j.

Assumption 3: The graph, $\mathcal{G}$, is strongly-connected.

We consider distributed iterative algorithms to solve Problem P1, where each agent is able to call a stochastic first-order oracle ($\mathcal{SFO}$). At iteration k and agent i, given $\mathbf{x}_k^i \in \mathbb{R}^p$ as the input, $\mathcal{SFO}$ returns a stochastic gradient in the form of $\mathbf{g}_i(\mathbf{x}_k^i, \xi_k^i) \in \mathbb{R}^p$, where $\xi_k^i \in \mathbb{R}^m$ are random vectors, $\forall k \geq 0, \forall i \in \mathcal{V}$. The stochastic gradients, $\mathbf{g}_i(\mathbf{x}_k^i, \xi_k^i)$, satisfy the following standard assumptions:

Assumption 4: The set of random vectors $\{\xi_k^i\}_{k \geq 0, i \in \mathcal{V}}$ are independent of each other, and

(1) $\mathbb{E}_{\xi_k^i}\left[\mathbf{g}_i(\mathbf{x}_k^i, \xi_k^i) | \mathbf{x}_k^i\right] = \nabla f_i(\mathbf{x}_k^i)$,

(2) $\mathbb{E}_{\xi_k^i}\left[\left\|\mathbf{g}_i(\mathbf{x}_k^i, \xi_k^i) - \nabla f_i(\mathbf{x}_k^i)\right\|_2^2 | \mathbf{x}_k^i\right] \leq \sigma^2$.

Assumption 4 is satisfied in many scenarios, for example, when the gradient noise, $\mathbf{g}_i(\mathbf{x}_k^i, \xi_i) - \nabla f_i(\mathbf{x}_k^i)$, is independent and identically distributed (i.i.d.) with zero-mean and finite second moment, while being independent of $\mathbf{x}_k^i$. However, Assumption 4 allows for general gradient noise processes dependent on agent i and the current iterate $\mathbf{x}_k^i$. Finally, we denote by $\mathcal{F}_k$ the σ-algebra generated by the set of random vectors $\{\xi_t^i\}_{0 \leq t \leq k-1, i \in \mathcal{V}}$.

A. The $\mathcal{S}$-$\mathcal{AB}$ algorithm

We now describe the proposed algorithm, $\mathcal{S}$-$\mathcal{AB}$, to solve Problem P1. Each agent i maintains two state vectors, $\mathbf{x}_k^i$ and $\mathbf{y}_k^i$, both in $\mathbb{R}^p$, where k is the number of iterations. The variable $\mathbf{x}_k^i$ is the estimate of the global minimizer $\mathbf{x}^*$, while $\mathbf{y}_k^i$ is the global gradient estimator. The $\mathcal{S}$-$\mathcal{AB}$ algorithm, initialized with arbitrary $\mathbf{x}_0^i$'s and with $\mathbf{y}_0^i = \mathbf{g}_i(\mathbf{x}_0^i, \xi_0^i)$, $\forall i \in \mathcal{V}$, is given by the following:

$$\mathbf{x}_{k+1}^i = \sum_{j=1}^n a_{ij} \mathbf{x}_k^j - \alpha \mathbf{y}_k^i, \tag{2a}$$

$$\mathbf{y}_{k+1}^i = \sum_{j=1}^n b_{ij} \mathbf{y}_k^j + \mathbf{g}_i(\mathbf{x}_{k+1}^i, \xi_{k+1}^i) - \mathbf{g}_i(\mathbf{x}_k^i, \xi_k^i), \tag{2b}$$

where the weight matrices $A = \{a_{ij}\}$ and $B = \{b_{ij}\}$ are row- and column-stochastic, respectively, and follow the graph topology, i.e., $a_{ij} > 0$ and $b_{ij} > 0$, iff $(i,j) \in \mathcal{E}$. We next write the algorithm in a compact vector form for the sake of analysis.

$$\mathbf{x}_{k+1} = \mathcal{A}\mathbf{x}_k - \alpha\mathbf{y}_k, \tag{3a}$$

$$\mathbf{y}_{k+1} = \mathcal{B}\mathbf{y}_k + \mathbf{g}(\mathbf{x}_{k+1}, \boldsymbol{\xi}_{k+1}) - \mathbf{g}(\mathbf{x}_k, \boldsymbol{\xi}_k), \tag{3b}$$

where we use the following notation:

$$\mathbf{x}_k \triangleq \begin{bmatrix} \mathbf{x}_k^1 \\ \vdots \\ \mathbf{x}_k^n \end{bmatrix}, \quad \mathbf{y}_k \triangleq \begin{bmatrix} \mathbf{y}_k^1 \\ \vdots \\ \mathbf{y}_k^n \end{bmatrix}, \quad \mathbf{g}(\mathbf{x}_k, \boldsymbol{\xi}_k) \triangleq \begin{bmatrix} \mathbf{g}_1(\mathbf{x}_k^1, \xi_k^1) \\ \vdots \\ \mathbf{g}_n(\mathbf{x}_k^n, \xi_k^n) \end{bmatrix},$$

and $\mathcal{A} = A \otimes I_p, \mathcal{B} = B \otimes I_p$.

Note that when the variance, σ, of the stochastic gradients is 0, we recover the $\mathcal{AB}$ or the push-pull algorithm proposed in [22], [23]. In the following, we assume $p = 1$ for the sake of simplicity. The analysis can be extended to the general case of $p > 1$ with the help of Kronecker products.

B. Auxiliary Results

We now provide some auxiliary results to aid the convergence analysis of $\mathcal{S}$-$\mathcal{AB}$. We first develop explicit norms regarding the contractions of the weight matrices, A and B. Since both A and B are primitive and stochastic, we use their non-$\mathbf{1}_n$ Perron vectors, $\boldsymbol{\pi}_r$ and $\boldsymbol{\pi}_c$, respectively, to define two weighted inner products as follows: $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$,

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\boldsymbol{\pi}_r} \triangleq \mathbf{x}^\top \mathrm{diag}(\boldsymbol{\pi}_r)\mathbf{y},$$

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\boldsymbol{\pi}_c} \triangleq \mathbf{x}^\top \mathrm{diag}(\boldsymbol{\pi}_c)^{-1}\mathbf{y}.$$

The above inner products are well-defined because the Perron vectors, $\boldsymbol{\pi}_r$ and $\boldsymbol{\pi}_c$, are positive and thus respectively induce a weighted Euclidean norm as follows: $\forall \mathbf{x} \in \mathbb{R}^n$,

$$\|\mathbf{x}\|_{\boldsymbol{\pi}_r} \triangleq \sqrt{[\boldsymbol{\pi}_r]_1 x_1^2 + \cdots + [\boldsymbol{\pi}_r]_n x_n^2} = \|\mathrm{diag}(\sqrt{\boldsymbol{\pi}_r})\mathbf{x}\|_2,$$

$$\|\mathbf{x}\|_{\boldsymbol{\pi}_c} \triangleq \sqrt{\frac{x_1^2}{[\boldsymbol{\pi}_c]_1} + \cdots + \frac{x_n^2}{[\boldsymbol{\pi}_c]_n}} = \|\mathrm{diag}(\sqrt{\boldsymbol{\pi}_c})^{-1}\mathbf{x}\|_2.$$

We denote $\vert\!\vert\!\vert \cdot \vert\!\vert\!\vert_{\boldsymbol{\pi}_r}$ and $\vert\!\vert\!\vert \cdot \vert\!\vert\!\vert_{\boldsymbol{\pi}_c}$ as the matrix norms induced by $\|\cdot\|_{\boldsymbol{\pi}_r}$ and $\|\cdot\|_{\boldsymbol{\pi}_c}$, respectively, i.e., $\forall X \in \mathbb{R}^{n \times n}$, see [34],

$$\vert\!\vert\!\vert X \vert\!\vert\!\vert_{\boldsymbol{\pi}_r} = \vert\!\vert\!\vert \mathrm{diag}(\sqrt{\boldsymbol{\pi}_r})X \mathrm{diag}(\sqrt{\boldsymbol{\pi}_r})^{-1} \vert\!\vert\!\vert_2, \tag{4}$$

$$\vert\!\vert\!\vert X \vert\!\vert\!\vert_{\boldsymbol{\pi}_c} = \vert\!\vert\!\vert \mathrm{diag}(\sqrt{\boldsymbol{\pi}_c})^{-1}X \mathrm{diag}(\sqrt{\boldsymbol{\pi}_c}) \vert\!\vert\!\vert_2. \tag{5}$$

It can be verified that the corresponding norm equivalence relationships between $\|\cdot\|_2$, $\|\cdot\|_{\boldsymbol{\pi}_r}$, and $\|\cdot\|_{\boldsymbol{\pi}_c}$ are given by

$$\|\cdot\|_{\boldsymbol{\pi}_r} \leq \overline{\boldsymbol{\pi}_r}^{-0.5}\|\cdot\|_2, \qquad\qquad \|\cdot\|_2 \leq \overline{\boldsymbol{\pi}_c}^{-0.5}\|\cdot\|_{\boldsymbol{\pi}_c},$$

$$\|\cdot\|_{\boldsymbol{\pi}_c} \leq \underline{\boldsymbol{\pi}_c}^{-0.5}\|\cdot\|_2, \qquad\qquad \|\cdot\|_2 \leq \underline{\boldsymbol{\pi}_r}^{-0.5}\|\cdot\|_{\boldsymbol{\pi}_r}.$$

We next establish the contraction of the A and B matrices with the help of the above arguments.

Lemma 1: Let Assumption 3 hold. Consider the weight matrices A, B in (3). We have: $\forall \mathbf{x} \in \mathbb{R}^n$,

$$\|A\mathbf{x} - A_\infty\mathbf{x}\|_{\boldsymbol{\pi}_r} \leq \sigma_A \|\mathbf{x} - A_\infty\mathbf{x}\|_{\boldsymbol{\pi}_r}, \tag{6}$$

$$\|B\mathbf{x} - B_\infty\mathbf{x}\|_{\boldsymbol{\pi}_c} \leq \sigma_B \|\mathbf{x} - B_\infty\mathbf{x}\|_{\boldsymbol{\pi}_c}, \tag{7}$$

with $\sigma_A \triangleq \vert\!\vert\!\vert A - A_\infty \vert\!\vert\!\vert_{\boldsymbol{\pi}_r} < 1$ and $\sigma_B \triangleq \vert\!\vert\!\vert B - B_\infty \vert\!\vert\!\vert_{\boldsymbol{\pi}_c} < 1$.

The proof of Lemma 1 is available in the Appendix. It can be further verified that

$$\sigma_A = \sigma_2\Big(\mathrm{diag}(\sqrt{\boldsymbol{\pi}_r})A\mathrm{diag}(\sqrt{\boldsymbol{\pi}_r})^{-1}\Big),$$

$$\sigma_B = \sigma_2\Big(\mathrm{diag}(\sqrt{\boldsymbol{\pi}_c})^{-1}B\mathrm{diag}(\sqrt{\boldsymbol{\pi}_c})\Big),$$

$$\vert\!\vert\!\vert A \vert\!\vert\!\vert_{\boldsymbol{\pi}_r} = \vert\!\vert\!\vert A_\infty \vert\!\vert\!\vert_{\boldsymbol{\pi}_r} = \vert\!\vert\!\vert I_n - A_\infty \vert\!\vert\!\vert_{\boldsymbol{\pi}_r} = 1,$$

$$\vert\!\vert\!\vert B \vert\!\vert\!\vert_{\boldsymbol{\pi}_c} = \vert\!\vert\!\vert B_\infty \vert\!\vert\!\vert_{\boldsymbol{\pi}_c} = \vert\!\vert\!\vert I_n - B_\infty \vert\!\vert\!\vert_{\boldsymbol{\pi}_c} = 1,$$

where $\sigma_2(\cdot)$ is the second largest singular value of a matrix.

In the following, Lemma 2 provides some simple results on the stochastic gradients, Lemma 3 uses the l-smoothness of the cost functions, while Lemmas 4 and 5 are standard in convex optimization and matrix analysis. To present these results, we define three quantities:

$$\overline{\mathbf{y}}_k \triangleq \frac{1}{n}\mathbf{1}_n^\top\mathbf{y}_k, \quad \mathbf{h}(\mathbf{x}_k) \triangleq \frac{1}{n}\mathbf{1}_n^\top\nabla\mathbf{f}(\mathbf{x}_k), \quad \widehat{\mathbf{x}}_k \triangleq \boldsymbol{\pi}_r^\top\mathbf{x}_k, \quad \overline{\mathbf{g}}(\mathbf{x}_k, \boldsymbol{\xi}_k) \triangleq \frac{1}{n}\mathbf{1}_n^\top\mathbf{g}(\mathbf{x}_k, \boldsymbol{\xi}_k),$$

where $\nabla\mathbf{f}(\mathbf{x}_k) \triangleq [\nabla f_1(\mathbf{x}_k^1)^\top, \ldots, \nabla f_n(\mathbf{x}_k^n)^\top]^\top$. The following statements use standard arguments and their formal proofs are omitted due to space limitations. Similar results can be found in [13], [22], [33].

Lemma 2: Consider the iterates $\{\mathbf{y}_k\}_{k \geq 0}$ generated by $\mathcal{S}\text{-}\mathcal{AB}$ in (3b) and let Assumptions 2 and 4 hold. Then the following hold, $\forall k \geq 0$:

(1) $\overline{\mathbf{y}}_k = \overline{\mathbf{g}}(\mathbf{x}_k, \boldsymbol{\xi}_k)$

(2) $\mathbb{E}[\overline{\mathbf{y}}_k | \mathcal{F}_k] = \mathbf{h}(\mathbf{x}_k)$

(3) $\mathbb{E}\left[\|\overline{\mathbf{y}}_k - \mathbf{h}(\mathbf{x}_k)\|_2^2 \,\middle|\, \mathcal{F}_k\right] \leq \frac{\sigma^2}{n}$

Lemma 3: Consider the iterates $\{\mathbf{x}_k\}_{k \geq 0}$ generated by the $\mathcal{S}\text{-}\mathcal{AB}$ algorithm in (2) and let Assumptions 2 hold. Then the following holds, $\forall k \geq 0$:

$$\|\mathbf{h}(\mathbf{x}_k) - \nabla F(\widehat{\mathbf{x}}_k)\|_2 \leq \frac{l}{\sqrt{n}}\|\mathbf{x}_k - \mathbf{1}_n \widehat{\mathbf{x}}_k\|_2.$$

Lemma 4 ([35]): Let Assumptions 1 and 2 hold. Recall that the global objective function, F, is μ-strongly-convex and l-smooth. If $0 < \alpha < \frac{1}{l}$, we have: $\forall \mathbf{x} \in \mathbb{R}^p$,

$$\|\mathbf{x} - \nabla F(\mathbf{x}) - \mathbf{x}^*\|_2 \leq (1 - \alpha\mu)\|\mathbf{x} - \mathbf{x}^*\|_2,$$

where $\mathbf{x}^*$ is the global minimizer of F.

Lemma 5 ([34]): Let $X \in \mathbb{R}^{n \times n}$ be non-negative and $\mathbf{x} \in \mathbb{R}^n$ be a positive vector. If $X\mathbf{x} < \omega\mathbf{x}$ with $\omega > 0$, then $\rho(X) < \omega$.

III. CONVERGENCE ANALYSIS

In this section, we analyze the $\mathcal{S}\text{-}\mathcal{AB}$ algorithm and establish its convergence properties for which we present Lemmas 6-9. The proofs for these lemmas are provided in the Appendix. First, in Lemma 6, we bound $\|\mathbf{y}_k\|_2^2$.

Lemma 6: Let Assumptions 1-4 hold. Then the iterates $\{\mathbf{y}_k\}_{k \geq 0}$ in (3) follow:

$$\mathbb{E}\left[\|\mathbf{y}_k\|_2^2 \,\middle|\, \mathcal{F}_k\right] \leq \frac{4n\|\boldsymbol{\pi}_c\|_2^2 l^2}{\underline{\boldsymbol{\pi}_r}}\|\mathbf{x}_k - \mathbf{1}_n\widehat{\mathbf{x}}_k\|_{\boldsymbol{\pi}_r}^2 + 4n^2\|\boldsymbol{\pi}_c\|_2^2 l^2 \|\widehat{\mathbf{x}}_k - \mathbf{x}^*\|_2^2$$
$$+ 4\overline{\boldsymbol{\pi}_c}\mathbb{E}\left[\|\mathbf{y}_k - B_\infty\mathbf{y}_k\|_{\boldsymbol{\pi}_c}^2 \,\middle|\, \mathcal{F}_k\right] + 4n\|\boldsymbol{\pi}_c\|_2^2\sigma^2. \tag{8}$$

Next in Lemmas 7-9, we bound the following three quantities in expectation, conditioned on the σ-algebra $\mathcal{F}_k$: (i) $\|\mathbf{x}_{k+1} - \mathbf{1}_n\widehat{\mathbf{x}}_{k+1}\|_{\boldsymbol{\pi}_r}^2$, the consensus error in the network; (ii) $\|\widehat{\mathbf{x}}_{k+1} - \mathbf{x}^*\|_2^2$, the optimality gap; and, (iii) $\|\mathbf{y}_{k+1} - B_\infty\mathbf{y}_{k+1}\|_{\boldsymbol{\pi}_c}^2$, the gradient tracking error. We then show that the norm of a vector composed of these three quantities converges linearly to a ball around the optimal when the step-size α is fixed and sufficiently small. The first lemma below is on the consensus error.

Lemma 7: Let Assumption 3 hold. Then the consensus error in the network follows: $\forall k \geq 0$,

$$\mathbb{E}\left[\|\mathbf{x}_{k+1} - \mathbf{1}_n\widehat{\mathbf{x}}_{k+1}\|_{\boldsymbol{\pi}_r}^2 \,\middle|\, \mathcal{F}_k\right] \leq \frac{1 + \sigma_A^2}{2}\|\mathbf{x}_k - \mathbf{1}_n\widehat{\mathbf{x}}_k\|_{\boldsymbol{\pi}_r}^2 + \frac{2\overline{\boldsymbol{\pi}_r}\alpha^2}{1 - \sigma_A^2}\mathbb{E}\left[\|\mathbf{y}_k\|_2^2 \,\middle|\, \mathcal{F}_k\right]. \tag{9}$$

The next lemma is on the optimality gap.

Lemma 8: Let Assumptions 1-4 hold. If $0 < \alpha < \frac{1}{n\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c l}$, the optimality gap in the network follows: $\forall k \geq 0$,

$$
\begin{aligned}
\mathbb{E}\left[\|\widehat{\mathbf{x}}_{k+1} - \mathbf{x}^*\|_2^2 \,|\, \mathcal{F}_k\right] \leq\ & \frac{3\alpha \boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c l^2}{\mu \underline{\boldsymbol{\pi}}_r} \|\mathbf{x}_k - \mathbf{1}_n \widehat{\mathbf{x}}_k\|_{\boldsymbol{\pi}_r}^2 \\
& + \left(1 - \frac{\mu n \boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c \alpha}{2}\right) \|\widehat{\mathbf{x}}_k - \mathbf{x}^*\|_2^2 \\
& + \frac{3\alpha \|\boldsymbol{\pi}_r\|_2^2 \overline{\boldsymbol{\pi}}_c}{\mu n \boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c} \mathbb{E}\left[\|\mathbf{y}_k - B_\infty \mathbf{y}_k\|_{\boldsymbol{\pi}_c}^2 \,|\, \mathcal{F}_k\right] \\
& + \frac{3\left(\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c\right)^2 n\alpha^2 \sigma^2}{2}.
\end{aligned}
\tag{10}
$$

Finally, we quantify the gradient tracking error.

Lemma 9: Let Assumptions 2-4 hold. The gradient tracking error follows:

$$
\begin{aligned}
\mathbb{E}\left[\|\mathbf{y}_{k+1} - B_\infty \mathbf{y}_{k+1}\|_{\boldsymbol{\pi}_c}^2 \,|\, \mathcal{F}_k\right] \leq\ & \frac{16 l^2}{\underline{\boldsymbol{\pi}}_r \, \underline{\boldsymbol{\pi}}_c (1 - \sigma_B^2)} \|\mathbf{x}_k - \mathbf{1}_n \widehat{\mathbf{x}}_k\|_{\boldsymbol{\pi}_r}^2 \\
& + \frac{1 + \sigma_B^2}{2} \mathbb{E}\left[\|\mathbf{y}_k - B_\infty \mathbf{y}_k\|_{\boldsymbol{\pi}_c}^2 \,|\, \mathcal{F}_k\right] \\
& + \frac{4 l^2 \alpha^2}{\underline{\boldsymbol{\pi}}_c (1 - \sigma_B^2)} \mathbb{E}\left[\|\mathbf{y}_k\|_2^2 \,|\, \mathcal{F}_k\right] + \frac{2nl\sigma^2\alpha}{\underline{\boldsymbol{\pi}}_c} + \frac{4n\sigma^2}{\underline{\boldsymbol{\pi}}_c}.
\end{aligned}
\tag{11}
$$

With the help of the above lemmas, we define a vector, $\mathbf{t}_k \in \mathbb{R}^3$, i.e.,

$$
\mathbf{t}_k = \begin{bmatrix} \mathbb{E}\left[\|\mathbf{x}_k - \mathbf{1}_n \widehat{\mathbf{x}}_k\|_{\boldsymbol{\pi}_r}^2\right] \\ \mathbb{E}\left[\|\widehat{\mathbf{x}}_k - \mathbf{x}^*\|_2^2\right] \\ \mathbb{E}\left[\|\mathbf{y}_k - B_\infty \mathbf{y}_k\|_{\boldsymbol{\pi}_c}^2\right]. \end{bmatrix}
$$

By substituting the bound on $\mathbb{E}[\|\mathbf{y}_k\|_2^2 \,|\, \mathcal{F}_k]$ from Lemma 6 in Lemmas 7-9, and taking the full expectation of both sides, it can be verified that $\mathbf{t}_k$ follows the dynamical system below.

$$
\begin{aligned}
\mathbf{t}_{k+1} \leq\ & \begin{bmatrix} \frac{1+\sigma_A^2}{2} + a_1\alpha^2 & a_2\alpha^2 & a_3\alpha^2 \\ a_4\alpha & 1 - a_5\alpha & a_6\alpha \\ a_7 + a_8\alpha^2 & a_9\alpha^2 & \frac{1+\sigma_B^2}{2} + a_{10}\alpha^2 \end{bmatrix} \mathbf{t}_k + \begin{bmatrix} b_1\alpha^2 \\ b_2\alpha^2 \\ b_3 + b_4\alpha + b_5\alpha^2 \end{bmatrix}, \\
\triangleq\ & G_\alpha \mathbf{t}_k + \mathbf{b}_\alpha,
\end{aligned}
\tag{12}
$$

where the constants are given by

$$
\begin{aligned}
a_1 &= \frac{8\overline{\boldsymbol{\pi}}_r n \|\boldsymbol{\pi}_c\|_2^2 l^2}{\underline{\boldsymbol{\pi}}_r (1 - \sigma_A^2)}, & a_8 &= \frac{16 n l^4 \|\boldsymbol{\pi}_c\|_2^2}{\underline{\boldsymbol{\pi}}_r \, \underline{\boldsymbol{\pi}}_c (1 - \sigma_B^2)}, \\
a_2 &= \frac{8\overline{\boldsymbol{\pi}}_r n^2 \|\boldsymbol{\pi}_c\|_2^2 l^2}{1 - \sigma_A^2}, & a_9 &= \frac{16 n^2 l^4 \|\boldsymbol{\pi}_c\|_2^2}{\underline{\boldsymbol{\pi}}_c (1 - \sigma_B^2)}, \\
a_3 &= \frac{8\overline{\boldsymbol{\pi}}_r \, \overline{\boldsymbol{\pi}}_c}{1 - \sigma_A^2}, & a_{10} &= \frac{16\overline{\boldsymbol{\pi}}_c l^2}{\underline{\boldsymbol{\pi}}_c (1 - \sigma_B^2)}, \\
a_4 &= \frac{3\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c l^2}{\mu \underline{\boldsymbol{\pi}}_r}, & b_1 &= \frac{8\overline{\boldsymbol{\pi}}_r n \|\boldsymbol{\pi}_c\|_2^2 \sigma^2}{1 - \sigma_A^2}, \\
a_5 &= \frac{\mu n \boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c}{2}, & b_2 &= \frac{3(\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c)^2 n\sigma^2}{2}, \\
a_6 &= \frac{3\|\boldsymbol{\pi}_r\|_2^2 \overline{\boldsymbol{\pi}}_c}{\mu n \boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c}, & b_3 &= \frac{4n\sigma^2}{\underline{\boldsymbol{\pi}}_c}, \\
a_7 &= \frac{16 l^2}{\underline{\boldsymbol{\pi}}_c \, \underline{\boldsymbol{\pi}}_r (1 - \sigma_B^2)}, & b_4 &= \frac{2nl\sigma^2}{\underline{\boldsymbol{\pi}}_c}, \\
b_5 &= \frac{16 n l^2 \|\boldsymbol{\pi}_c\|_2^2 \sigma^2}{\underline{\boldsymbol{\pi}}_r \, \underline{\boldsymbol{\pi}}_c (1 - \sigma_B^2)}. &
\end{aligned}
$$

IV. MAIN RESULT

In this section, we analyze the inequality on $\mathbf{t}_k$ to establish the convergence of $\mathcal{S}\text{-}\mathcal{AB}$.

Theorem 1: Consider the $\mathcal{S}\text{-}\mathcal{AB}$ algorithm in (2) and let Assumptions 1-4 hold. Suppose the step-size α satisfies the following the condition:

$$0 < \alpha < \min \left\{ \frac{1}{ln\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c}, \frac{\left(1 - \sigma_A^2\right)\sqrt{\left(1 - \sigma_B^2\right)}\left(\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c\right)}{l\kappa \left\|\boldsymbol{\pi}_r\right\|_2 \left\|\boldsymbol{\pi}_c\right\|_2 \sqrt{384 h_r h_c (n \left\|\boldsymbol{\pi}_c\right\|_2^2 + 64)}}, \frac{\left(1 - \sigma_B^2\right)\left(\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c\right)}{l\kappa \left\|\boldsymbol{\pi}_r\right\|_2 \left\|\boldsymbol{\pi}_c\right\|_2 \sqrt{24\left(n \left\|\boldsymbol{\pi}_c\right\|_2^2 + 48 h_c\right)}} \right\},$$

where $h_r = \overline{\boldsymbol{\pi}_r}/\underline{\boldsymbol{\pi}_r}$, $h_c = \overline{\boldsymbol{\pi}_c}/\underline{\boldsymbol{\pi}_c}$ and $\kappa = l/\mu$. Then, $\rho(G_\alpha) < 1$, the vector $(I_3 - G_\alpha)^{-1}\mathbf{b}_\alpha$ has non-negative components, and we have that

$$\limsup_{k \to \infty} \mathbf{t}_k \leq (I_3 - G_\alpha)^{-1}\mathbf{b}_\alpha$$

$$\limsup_{k \to \infty} \mathbb{E}\left[\|\mathbf{x}_k - \mathbf{1}_n \widehat{\mathbf{x}}_k\|_{\boldsymbol{\pi}_r}^2\right] \leq \left[(I_3 - G_\alpha)^{-1}\mathbf{b}_\alpha\right]_1$$

$$\limsup_{k \to \infty} \mathbb{E}\left[\|\widehat{\mathbf{x}}_k - \mathbf{x}^*\|_2^2\right] \leq \left[(I_3 - G_\alpha)^{-1}\mathbf{b}_\alpha\right]_2,$$

where the above convergence is geometric with exponent $\rho(G_\alpha)$.

Proof: The goal is to find the range of α such that $\rho(G_\alpha) < 1$. In the light of Lemma 5, it suffices to solve for the range of α such that $G_\alpha \boldsymbol{\delta} < \boldsymbol{\delta}$ holds for some positive vector $\boldsymbol{\delta} = [\delta_1, \delta_2, \delta_3]^\top$. We now expand this element-wise matrix inequality as follows:

$$\left(\tfrac{1+\sigma_A^2}{2} + a_1\alpha^2\right)\delta_1 + a_2\alpha^2\delta_2 + a_3\alpha^2\delta_3 < \delta_1$$

$$a_4\alpha\delta_1 + (1 - a_5\alpha)\delta_2 + a_6\alpha\delta_3 < \delta_2$$

$$\left(a_7 + a_8\alpha^2\right)\delta_1 + a_9\alpha^2\delta_2 + \left(\tfrac{1+\sigma_B^2}{2} + a_{10}\alpha^2\right)\delta_3 < \delta_3$$

which can be reformulated as:

$$(a_1\delta_1 + a_2\delta_2 + a_3\delta_3)\,\alpha^2 < \tfrac{1-\sigma_A^2}{2}\delta_1 \tag{13}$$

$$a_4\delta_1\alpha - a_5\delta_2\alpha + a_6\delta_3\alpha < 0 \tag{14}$$

$$(a_8\delta_1 + a_9\delta_2 + a_{10}\delta_3)\,\alpha^2 < \tfrac{1-\sigma_B^2}{2}\delta_3 - a_7\delta_1. \tag{15}$$

We first note that in order for (14) to hold, $\delta_1, \delta_2, \delta_3$ suffice to satisfy:

$$\delta_2 > \frac{a_4\delta_1 + a_6\delta_3}{a_5} \tag{16}$$

Next, for the right handside of (15) to be positive, δ_1, δ_3 suffice to satisfy:

$$0 < \delta_1 < \frac{1 - \sigma_B^2}{2a_7}\delta_3 \tag{17}$$

In order to obtain an explicit upper bound on the step-size, $\overline{\alpha}$, such that (13)-(15) hold when $0 < \alpha < \overline{\alpha}$, we set $\delta_1, \delta_2, \delta_3$ as the following which satisfies (16) and (17):

$$\delta_1 = 1, \delta_2 = \frac{2}{a_5}\left(a_4 + \frac{4a_6 a_7}{1 - \sigma_B^2}\right), \delta_3 = \frac{4a_7}{1 - \sigma_B^2}. \tag{18}$$

Then we plug the values of $\delta_1, \delta_2, \delta_3$ in (18) into (13) and (14) to solve an upper bound on the step-size, α. From (13), we get:

$$\left(a_1 + \frac{2a_2}{a_5}\left(a_4 + \frac{4a_6 a_7}{1 - \sigma_B^2}\right) + \frac{4a_3 a_7}{1 - \sigma_B^2}\right)\alpha^2 < \frac{1 - \sigma_A^2}{2} \tag{19}$$

One can further verify that:

$$a_1 + \frac{2a_2}{a_5}\left(a_4 + \frac{4a_6 a_7}{1 - \sigma_B^2}\right) + \frac{4a_3 a_7}{1 - \sigma_B^2} < \frac{192 h_r h_c \|\boldsymbol{\pi}_r\|_2^2 \|\boldsymbol{\pi}_c\|_2^2 l^2 \kappa^2 (64 + n \|\boldsymbol{\pi}_c\|_2^2)}{\left(1 - \sigma_A^2\right)\left(1 - \sigma_B^2\right)\left(\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c\right)^2}$$

We use the above inequality in (19) to obtain an upper bound on α:

$$\alpha < \frac{\left(1 - \sigma_A^2\right)\left(\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c\right)}{l\kappa \|\boldsymbol{\pi}_r\|_2 \|\boldsymbol{\pi}_c\|_2}\sqrt{\frac{\left(1 - \sigma_B^2\right)}{384 h_r h_c (64 + n \|\boldsymbol{\pi}_c\|_2^2)}} \tag{20}$$

Similarly, from (15), we get:

$$\left(\frac{a_8}{a_7} + \frac{2a_9}{a_5 a_7}\left(a_4 + \frac{4a_6 a_7}{1 - \sigma_B^2}\right) + \frac{4a_{10}}{1 - \sigma_B^2}\right)\alpha^2 < 1 \tag{21}$$

One can then verify that:

$$\frac{a_8}{a_7} + \frac{2a_9}{a_5 a_7}\left(a_4 + \frac{4a_6 a_7}{1 - \sigma_B^2}\right) + \frac{4a_{10}}{1 - \sigma_B^2} < \frac{24 l^2 \kappa^2 \|\boldsymbol{\pi}_r\|_2^2 \|\boldsymbol{\pi}_c\|_2^2 \left(n \|\boldsymbol{\pi}_c\|_2^2 + 48 h_c\right)}{\left(1 - \sigma_B^2\right)^2 \left(\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c\right)^2}$$

Using the above inequality in (21) obtains another upper bound on α:

$$\alpha < \frac{\left(1 - \sigma_B^2\right)\left(\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c\right)}{l\kappa \|\boldsymbol{\pi}_r\|_2 \|\boldsymbol{\pi}_c\|_2 \sqrt{24\left(n \|\boldsymbol{\pi}_c\|_2^2 + 48 h_c\right)}} \tag{22}$$

We complete the proof by combining (20), (22) and the requirement from Lemma 8 that $0 < \alpha < \frac{1}{n\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c l}$. $\blacksquare$
It is important to note that the error bounds in Theorem 1 go to zero as the step-size gets smaller and the variance on the gradient noise decreases.

V. NUMERICAL EXPERIMENTS

In this section, we illustrate the $\mathcal{S}$-$\mathcal{AB}$ algorithm and its convergence properties. We demonstrate the results on a directed graph generated using nearest neighbor rules with $n = 20$ agents. The particular graph for the experiments is shown in Fig. 1 (left) to provide a sense of connectivity. We choose a logistic regression problem to classify around $12,000$ images of two digits, 7 and 8, labeled as $y_{ij} = +1$ or -1, from the MNIST dataset [36]. Each image, $\mathbf{c}_{ij}$, is a 785-dimensional vector and the total images are divided among the agents such that each agent has $m_i = 600$ images. Because privacy and communication restrictions, the agents do not share their local batches (local training images) with each other. In order to use the entire data set for training, the network of agents cooperatively solves the following distributed logistic regression problem:

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

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

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

We show the performance of this classification problem over centralized and distributed methods. Centralized gradient descent (CGD) uses the entire batch, i.e., it computes $12,000$ gradients at each iteration, whereas centralized stochastic gradient descent (C-SGD) uses only one data point at each iteration that is uniformly sampled from the entire batch. For the distributed algorithms, we show the performance of non-stochastic $\mathcal{AB}$, where each agent uses its entire local batch, i.e., 600 labeled data points. Whereas, for the implementation of $\mathcal{S}\text{-}\mathcal{AB}$, each agent uniformly chooses one data point from its local batch. For testing, we use 2000 additional images that were not used for training. The residuals are shown in Fig. 1 (right) while the training and testing accuracy is shown in Fig. 2. In the performance figures, the horizontal axis represents the number of epochs where each epoch represents computations on the entire batch. Clearly, $\mathcal{S}\text{-}\mathcal{AB}$ has a better performance when compared to $\mathcal{AB}$ in [22] as expected from the performance of their centralized counterparts, C-SGD and CGD.

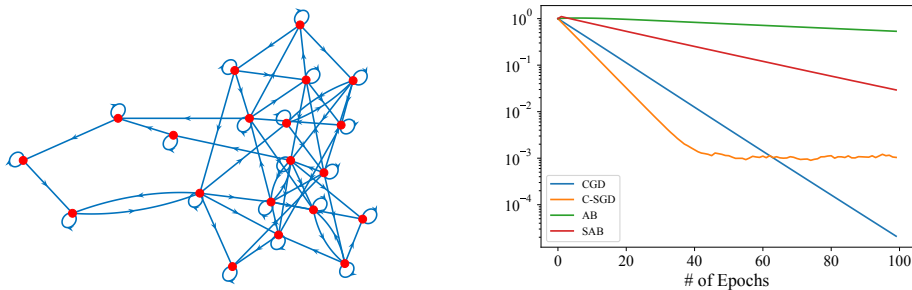

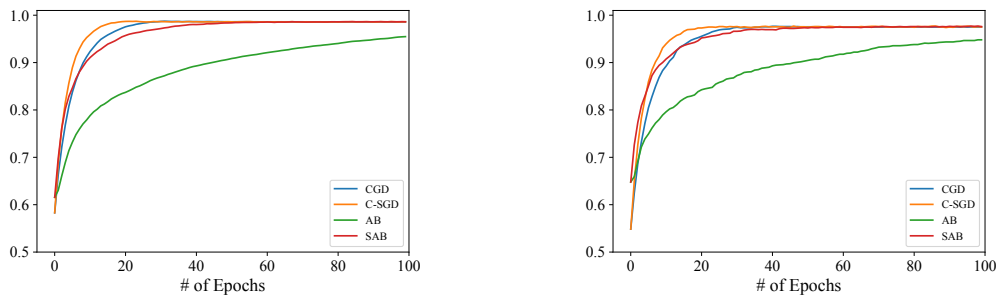Fig. 1.   (Left) Strongly-connected directed graph. (Right) Residuals.

Fig. 2.   (Left) Training accuracy. (Right) Test accuracy.

VI. Conclusions

In this paper, we have presented a stochastic gradient descent algorithm, $\mathcal{S}\text{-}\mathcal{AB}$, over arbitrary strongly-connected graphs. In this setup, the data is distributed over agents and each agent uniformly samples a data

point (from its local batch) at each iteration of the algorithm to implement the stochastic $\mathcal{S}$-$\mathcal{AB}$ algorithm. To cope with general directed communication graphs and potential lack of doubly-stochastic weights, $\mathcal{S}$-$\mathcal{AB}$ employs a two-phase update with row- and column-stochastic weights. We have shown that under a sufficiently small constant step-size, $\mathcal{S}$-$\mathcal{AB}$ converges linearly to a neighborhood of the global minimizer when the local cost functions are smooth and strongly-convex. We have presented numerical simulations based on real-world datasets to illustrate the theoretical results.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*, 2017, pp. 1273–1282.

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

[3] A. Daneshmand, F. Facchinei, V. Kungurtsev, and G. Scutari, "Hybrid random/deterministic parallel algorithms for convex and nonconvex big data optimization," *IEEE Trans. on Signal Processing*, vol. 63, no. 15, pp. 3914–3929, 2015.

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

[5] F. Bullo, J. Cortés, and S. Martinez, *Distributed control of robotic networks: A mathematical approach to motion coordination algorithms*, Princeton University Press, 2009.

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

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

[8] I. Lobel, A. Ozdaglar, and D. Feijer, "Distributed multi-agent optimization with state-dependent communication," *Mathematical Programming*, vol. 129, no. 2, pp. 255–284, 2011.

[9] J. C. Duchi, A. Agarwal, and M. J. Wainwright, "Dual averaging for distributed optimization: Convergence analysis and network scaling," *IEEE Trans. on Automatic Control*, vol. 57, no. 3, pp. 592–606, Mar. 2012.

[10] S. S. Ram, A. Nedić, 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.

[11] A. Nedić and A. Olshevsky, "Stochastic gradient-push for strongly convex functions on time-varying directed graphs," *IEEE Trans. on Automatic Control*, vol. 61, no. 12, pp. 3936–3947, Dec. 2016.

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

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

[14] A. Nedić, A. Olshevsky, and W. Shi, "Achieving geometric convergence for distributed optimization over time-varying graphs," *SIAM Journal of Optimization*, vol. 27, no. 4, pp. 2597–2633, Dec. 2017.

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

[16] 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.

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

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

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

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

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

[22] R. Xin and U. A. Khan, "A linear algorithm for optimization over directed graphs with geometric convergence," *IEEE Control Systems Letters*, vol. 2, no. 3, pp. 325–330, Jul. 2018.

[23] S. Pu, W. Shi, J. Xu, and A. Nedić, "A push-pull gradient method for distributed optimization in networks," in *57th IEEE Annual Conference on Decision and Control*, Dec. 2018.

[24] R. Xin and U. A. Khan, "Distributed heavy-ball: A generalization and acceleration of first-order methods with gradient tracking," *arXiv preprint arXiv:1808.02942*, 2018.

[25] S. Pu, W. Shi, J. Xu, and A. Nedić, "Push-pull gradient methods for distributed optimization in networks," *https://arxiv.org/abs/1810.06653*, 2018.

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

[27] A. Daneshmand, G. Scutari, and V. Kungurtsev, "Second-order guarantees of distributed gradient algorithms," *arXiv preprint arXiv:1809.08694*, 2018.

[28] D. Yuan, Y. Hong, D. W. C. Ho, and G. Jiang, "Optimal distributed stochastic mirror descent for strongly convex optimization," *Automatica*, vol. 90, pp. 196–203, Apr. 2018.

[29] N. Denizcan Vanli, Muhammed O. Sayin, and Suleyman S. Kozat, "Stochastic subgradient algorithms for strongly convex optimization over distributed networks," *IEEE Trans. on Network Science and Engineering*, vol. 4, no. 4, pp. 248–260, Oct. 2017.

[30] D. Jakovetić, D. Bajović, A. K. Sahu, and S. Kar, "Convergence rates for distributed stochastic optimization over random networks," in *IEEE Conference on Decision and Control*, Dec. 2018, pp. 4238–4245.

[31] A. K. Sahu, D. Jakovetić, D. Bajović, and S. Kar, "Distributed zeroth order optimization over random networks: A Kiefer-Wolfowitz stochastic approximation approach," in *Conference on Decision and Control*, Dec. 2018, pp. 4951–4958.

[32] D. Jakovetic, "A unification and generalization of exact distributed first order methods," *IEEE Trans. on Signal and Information Processing over Networks*, 2018.

[33] S. Pu and A. Nedić, "A distributed stochastic gradient tracking method," in *2018 IEEE Conference on Decision and Control (CDC)*, Dec. 2018, pp. 963–968.

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

[35] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*, vol. 87, Springer Science & Business Media, 2013.

[36] Y. LeCun, C. Cortes, and C. Burges, "Mnist handwritten digit database," *AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist*, vol. 2, pp. 18, 2010.

APPENDIX

PROOF OF LEMMA 1

Proof: We start with the proof of (6). Note that $A\mathbf{x} - A_\infty\mathbf{x} = (A - A_\infty)(\mathbf{x} - A_\infty\mathbf{x})$ that leads to

$$\|A\mathbf{x} - A_\infty\mathbf{x}\|_{\boldsymbol{\pi}_r} \leq \||A - A_\infty\||_{\boldsymbol{\pi}_r} \|\mathbf{x} - A_\infty\mathbf{x}\|_{\boldsymbol{\pi}_r}.$$

By the definition of $\left\|\!\left\|\cdot\right\|\!\right\|_{\boldsymbol{\pi}_r}$ in (4), we have

$$\left\|\!\left\| A - A_\infty \right\|\!\right\|_{\boldsymbol{\pi}_r} = \left\|\!\left\| \mathrm{diag}(\sqrt{\boldsymbol{\pi}_r})(A - A_\infty)\mathrm{diag}(\sqrt{\boldsymbol{\pi}_r})^{-1} \right\|\!\right\|_2 \triangleq \sqrt{\lambda_{\max}(J)},$$

where $J = \mathrm{diag}(\sqrt{\boldsymbol{\pi}_r})^{-1}(A - A_\infty)^\top \mathrm{diag}(\boldsymbol{\pi}_r)(A - A_\infty)\mathrm{diag}(\sqrt{\boldsymbol{\pi}_r})^{-1}$ and $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue of the matrix. What we need to show is that $\rho(J) < 1$. Expanding J, we get

$$J = \mathrm{diag}(\sqrt{\boldsymbol{\pi}_r})^{-1}A^\top \mathrm{diag}(\boldsymbol{\pi}_r)A\mathrm{diag}(\sqrt{\boldsymbol{\pi}_r})^{-1} - \mathrm{diag}(\sqrt{\boldsymbol{\pi}_r})^{-1}A_\infty^\top \mathrm{diag}(\boldsymbol{\pi}_r)A\mathrm{diag}(\sqrt{\boldsymbol{\pi}_r})^{-1}$$
$$- \mathrm{diag}(\sqrt{\boldsymbol{\pi}_r})^{-1}A^\top \mathrm{diag}(\boldsymbol{\pi}_r)A_\infty \mathrm{diag}(\sqrt{\boldsymbol{\pi}_r})^{-1} + \mathrm{diag}(\sqrt{\boldsymbol{\pi}_r})^{-1}A_\infty^\top \mathrm{diag}(\boldsymbol{\pi}_r)A_\infty \mathrm{diag}(\sqrt{\boldsymbol{\pi}_r})^{-1},$$
$$\triangleq J_1 - J_2 - J_3 + J_4.$$

With the fact that $A_\infty = \mathbf{1}_n \boldsymbol{\pi}_r^\top$, it can be verified that $J_2 = J_3 = J_4 = \sqrt{\boldsymbol{\pi}_r}\sqrt{\boldsymbol{\pi}_r}^\top$ and thus $J = J_1 - \sqrt{\boldsymbol{\pi}_r}\sqrt{\boldsymbol{\pi}_r}^\top$. Furthermore, $J_1\sqrt{\boldsymbol{\pi}_r} = \sqrt{\boldsymbol{\pi}_r}$, and $\sqrt{\boldsymbol{\pi}_r}^\top J_1 = \sqrt{\boldsymbol{\pi}_r}^\top$. Since J_1 is primitive, by Perron-Frobenius theorem [34], we have $\rho(J) = \rho(J_1 - \sqrt{\boldsymbol{\pi}_r}\sqrt{\boldsymbol{\pi}_r}^\top) < 1$ and thus

$$\sigma_A \triangleq \left\|\!\left\| A - A_\infty \right\|\!\right\|_{\boldsymbol{\pi}_r} = \sqrt{\rho(J_1 - \sqrt{\boldsymbol{\pi}_r}\sqrt{\boldsymbol{\pi}_r}^\top)} < 1.$$

To prove (7), we note that $B\mathbf{x} - B_\infty \mathbf{x} = (B - B_\infty)(\mathbf{x} - B_\infty \mathbf{x})$ and we have the following:

$$\left\| B\mathbf{x} - B_\infty \mathbf{x} \right\|_{\boldsymbol{\pi}_c} \leq \left\|\!\left\| B - B_\infty \right\|\!\right\|_{\boldsymbol{\pi}_c} \left\| \mathbf{x} - B_\infty \mathbf{x} \right\|_{\boldsymbol{\pi}_c}.$$

Next we show that $\left\|\!\left\| B - B_\infty \right\|\!\right\|_{\boldsymbol{\pi}_c} < 1$. By the definition of $\left\|\!\left\|\cdot\right\|\!\right\|_{\boldsymbol{\pi}_c}$ in (5), we have the following:

$$\left\|\!\left\| B - B_\infty \right\|\!\right\|_{\boldsymbol{\pi}_c} = \left\|\!\left\| \mathrm{diag}(\sqrt{\boldsymbol{\pi}_c})^{-1}(B - B_\infty)\mathrm{diag}(\sqrt{\boldsymbol{\pi}_c}) \right\|\!\right\|_2 \triangleq \sqrt{\lambda_{\max}(H)},$$

where $H = \mathrm{diag}(\sqrt{\boldsymbol{\pi}_c})(B - B_\infty)^\top \mathrm{diag}(\boldsymbol{\pi}_c)^{-1}(B - B_\infty)\mathrm{diag}(\sqrt{\boldsymbol{\pi}_c})$. Next we show that $\rho(H) < 1$. We expand the expression for H as below:

$$H = \mathrm{diag}(\sqrt{\boldsymbol{\pi}_c})B^\top \mathrm{diag}(\boldsymbol{\pi}_c)^{-1}B\mathrm{diag}(\sqrt{\boldsymbol{\pi}_c}) - \mathrm{diag}(\sqrt{\boldsymbol{\pi}_c})B_\infty^\top \mathrm{diag}(\boldsymbol{\pi}_c)^{-1}B\mathrm{diag}(\sqrt{\boldsymbol{\pi}_c})$$
$$- \mathrm{diag}(\sqrt{\boldsymbol{\pi}_c})B^\top \mathrm{diag}(\boldsymbol{\pi}_c)^{-1}B_\infty \mathrm{diag}(\sqrt{\boldsymbol{\pi}_c}) + \mathrm{diag}(\sqrt{\boldsymbol{\pi}_c})B_\infty^\top \mathrm{diag}(\boldsymbol{\pi}_c)^{-1}B_\infty \mathrm{diag}(\sqrt{\boldsymbol{\pi}_c}),$$
$$\triangleq H_1 - H_2 - H_3 + H_4.$$

With the fact that $B_\infty = \boldsymbol{\pi}_c \mathbf{1}_n^\top$, one can verify that $H_2 = H_3 = H_4 = \sqrt{\boldsymbol{\pi}_c}\sqrt{\boldsymbol{\pi}_c}^\top$ and thus $H = H_1 - \sqrt{\boldsymbol{\pi}_c}\sqrt{\boldsymbol{\pi}_c}^\top$. Since H_1 is primitive, by Perron-Frobenius theorem [34], we have that

$$\sigma_B \triangleq \left\|\!\left\| B - B_\infty \right\|\!\right\|_{\boldsymbol{\pi}_c} = \sqrt{\rho(H_1 - \sqrt{\boldsymbol{\pi}_c}\sqrt{\boldsymbol{\pi}_c}^\top)} < 1,$$

which completes the proof. ∎

PROOF OF LEMMA 6

Proof: Recall that $B_\infty = \boldsymbol{\pi}_c \mathbf{1}_n^\top$. We have the following hold:

$$\|\mathbf{y}_k\|_2 \leq \overline{\boldsymbol{\pi}_c}^{-0.5}\|\mathbf{y}_k - \boldsymbol{\pi}_c \mathbf{1}_n^\top \mathbf{y}_k\|_{\boldsymbol{\pi}_c} + \|\boldsymbol{\pi}_c \mathbf{1}_n^\top \mathbf{y}_k\|_2$$

$$= \overline{\boldsymbol{\pi}_c}^{-0.5}\|\mathbf{y}_k - B_\infty \mathbf{y}_k\|_{\boldsymbol{\pi}_c} + \|\boldsymbol{\pi}_c\|_2\|\mathbf{1}_n^\top \mathbf{y}_k\|_2$$

$$\leq \overline{\boldsymbol{\pi}_c}^{-0.5}\|\mathbf{y}_k - B_\infty \mathbf{y}_k\|_{\boldsymbol{\pi}_c} + n\|\boldsymbol{\pi}_c\|_2\|\overline{\mathbf{y}}_k - \mathbf{h}(\mathbf{x}_k)\|_2 + n\|\boldsymbol{\pi}_c\|_2\|\mathbf{h}(\mathbf{x}_k) - \nabla F(\widehat{\mathbf{x}}_k)\|_2$$

$$+ n\|\boldsymbol{\pi}_c\|_2\|\nabla F(\widehat{\mathbf{x}}_k) - \nabla F(\mathbf{x}^*)\|_2$$

$$\leq \overline{\boldsymbol{\pi}_c}^{-0.5}\|\mathbf{y}_k - B_\infty \mathbf{y}_k\|_{\boldsymbol{\pi}_c} + n\|\boldsymbol{\pi}_c\|_2\|\overline{\mathbf{y}}_k - \mathbf{h}(\mathbf{x}_k)\|_2 + \frac{\sqrt{n}\|\boldsymbol{\pi}_c\|_2 l}{\sqrt{\underline{\boldsymbol{\pi}_r}}}\|\mathbf{x}_k - \mathbf{1}_n\widehat{\mathbf{x}}_k\|_{\boldsymbol{\pi}_r}$$

$$+ n\|\boldsymbol{\pi}_c\|_2 l \|\widehat{\mathbf{x}}_k - \mathbf{x}^*\|_2,$$

where in the last inequality we used Lemma 3 and norm-equivalence. Squaring the above, and using the basic inequality $2ab \leq a^2 + b^2$, we get

$$\|\mathbf{y}_k\|_2^2 \leq 4\overline{\boldsymbol{\pi}_c}\|\mathbf{y}_k - B_\infty \mathbf{y}_k\|_{\boldsymbol{\pi}_c}^2 + 4n^2\|\boldsymbol{\pi}_c\|_2^2\|\overline{\mathbf{y}}_k - \mathbf{h}(\mathbf{x}_k)\|_2^2$$

$$+ \frac{4n\|\boldsymbol{\pi}_c\|_2^2 l^2}{\underline{\boldsymbol{\pi}_r}}\|\mathbf{x}_k - \mathbf{1}_n\widehat{\mathbf{x}}_k\|_{\boldsymbol{\pi}_r}^2 + 4n^2\|\boldsymbol{\pi}_c\|_2^2 l^2\|\widehat{\mathbf{x}}_k - \mathbf{x}^*\|_2^2.$$

Taking the expectation on both sides given $\mathcal{F}_k$ and using (3) in Lemma 2 completes the proof. ∎

PROOF OF LEMMA 7

Proof: Recall that $A_\infty = \mathbf{1}_n \boldsymbol{\pi}_r^\top$ and $\widehat{\mathbf{x}}_k \triangleq \boldsymbol{\pi}_r^\top \mathbf{x}_k$. Using (3a) and Lemma 1, We have the following hold:

$$\|\mathbf{x}_{k+1} - \mathbf{1}_n\widehat{\mathbf{x}}_{k+1}\|_{\boldsymbol{\pi}_r}^2$$

$$= \|A\mathbf{x}_k - \alpha\mathbf{y}_k - A_\infty(A\mathbf{x}_k - \alpha\mathbf{y}_k)\|_{\boldsymbol{\pi}_r}^2$$

$$= \|A\mathbf{x}_k - A_\infty\mathbf{x}_k - \alpha(I_n - A_\infty)\mathbf{y}_k\|_{\boldsymbol{\pi}_r}^2$$

$$= \|A\mathbf{x}_k - A_\infty\mathbf{x}_k\|_{\boldsymbol{\pi}_r}^2 + \alpha^2\|(I_n - A_\infty)\mathbf{y}_k\|_{\boldsymbol{\pi}_r}^2 - 2\langle A\mathbf{x}_k - A_\infty\mathbf{x}_k, \alpha(I_n - A_\infty)\mathbf{y}_k\rangle_{\boldsymbol{\pi}_r}$$

$$\leq \sigma_A^2\|\mathbf{x}_k - A_\infty\mathbf{x}_k\|_{\boldsymbol{\pi}_r}^2 + \alpha^2\||I_n - A_\infty\||_{\boldsymbol{\pi}_r}^2\|\mathbf{y}_k\|_{\boldsymbol{\pi}_r}^2 + 2\alpha\sigma_A\|\mathbf{x}_k - A_\infty\mathbf{x}_k\|_{\boldsymbol{\pi}_r}\||I_n - A_\infty\||_{\boldsymbol{\pi}_r}\|\mathbf{y}_k\|_{\boldsymbol{\pi}_r}$$

$$\leq \sigma_A^2\|\mathbf{x}_k - A_\infty\mathbf{x}_k\|_{\boldsymbol{\pi}_r}^2 + \alpha^2\|\mathbf{y}_k\|_{\boldsymbol{\pi}_r}^2 + \alpha\sigma_A\left(\frac{1 - \sigma_A^2}{2\alpha\sigma_A}\|\mathbf{x}_k - A_\infty\mathbf{x}_k\|_{\boldsymbol{\pi}_r}^2 + \frac{2\alpha\sigma_A}{1 - \sigma_A^2}\|\mathbf{y}_k\|_{\boldsymbol{\pi}_r}^2\right)$$

$$\leq \frac{1 + \sigma_A^2}{2}\|\mathbf{x}_k - A_\infty\mathbf{x}_k\|_{\boldsymbol{\pi}_r}^2 + \frac{\alpha^2(1 + \sigma_A^2)\overline{\boldsymbol{\pi}_r}}{1 - \sigma_A^2}\|\mathbf{y}_k\|_2^2.$$

where the second last inequality uses Young's inequality and the fact that $\||I_n - A_\infty\||_{\boldsymbol{\pi}_r} = 1$. Taking the expectation on both sides given $\mathcal{F}_k$ completes the proof. ∎

PROOF OF LEMMA 8

Proof: We start by multiplying both sides of (3a) with $\boldsymbol{\pi}_r^\top$ to obtain as in [22], [23]:

$$\widehat{\mathbf{x}}_{k+1} - \mathbf{x}^* = \widehat{\mathbf{x}}_k - \mathbf{x}^* - \alpha\boldsymbol{\pi}_r^\top(\mathbf{y}_k - \boldsymbol{\pi}_c\mathbf{1}_n^\top \mathbf{y}_k + \boldsymbol{\pi}_c\mathbf{1}_n^\top \mathbf{y}_k)$$

$$= \widehat{\mathbf{x}}_k - \mathbf{x}^* - n\alpha\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c \overline{\mathbf{y}}_k - \alpha\boldsymbol{\pi}_r^\top(\mathbf{y}_k - B_\infty \mathbf{y}_k).$$

Taking norms and squaring both sides leads to

$$\|\widehat{\mathbf{x}}_{k+1} - \mathbf{x}^*\|_2^2$$

$$= \left\| \widehat{\mathbf{x}}_k - \mathbf{x}^* - n\alpha\boldsymbol{\pi}_r^\top\boldsymbol{\pi}_c\overline{\mathbf{y}}_k - \alpha\boldsymbol{\pi}_r^\top(\mathbf{y}_k - B_\infty\mathbf{y}_k) \right\|_2^2$$

$$= \left\| \widehat{\mathbf{x}}_k - \mathbf{x}^* - n\alpha\boldsymbol{\pi}_r^\top\boldsymbol{\pi}_c\overline{\mathbf{y}}_k \right\|_2^2 - 2\left\langle \widehat{\mathbf{x}}_k - \mathbf{x}^* - n\alpha\boldsymbol{\pi}_r^\top\boldsymbol{\pi}_c\overline{\mathbf{y}}_k, \alpha\boldsymbol{\pi}_r^\top(\mathbf{y}_k - B_\infty\mathbf{y}_k) \right\rangle + \alpha^2 \left\| \boldsymbol{\pi}_r^\top(\mathbf{y}_k - B_\infty\mathbf{y}_k) \right\|_2^2$$

$$\leq \left\| \widehat{\mathbf{x}}_k - \mathbf{x}^* - n\alpha\boldsymbol{\pi}_r^\top\boldsymbol{\pi}_c\overline{\mathbf{y}}_k \right\|_2^2 - 2\left\langle \widehat{\mathbf{x}}_k - \mathbf{x}^* - n\alpha\boldsymbol{\pi}_r^\top\boldsymbol{\pi}_c\overline{\mathbf{y}}_k, \alpha\boldsymbol{\pi}_r^\top(\mathbf{y}_k - B_\infty\mathbf{y}_k) \right\rangle + \alpha^2\|\boldsymbol{\pi}_r\|_2^2\overline{\boldsymbol{\pi}_c}\|\mathbf{y}_k - B_\infty\mathbf{y}_k\|_{\boldsymbol{\pi}_c}^2$$

$$\triangleq r_1 + r_2 + \alpha^2\|\boldsymbol{\pi}_r\|_2^2\overline{\boldsymbol{\pi}_c}\|\mathbf{y}_k - B_\infty\mathbf{y}_k\|_{\boldsymbol{\pi}_c}^2.$$

Taking the conditional expectation on bothsides given $\mathcal{F}_k$, we obtain:

$$\mathbb{E}\left[\|\widehat{\mathbf{x}}_{k+1} - \mathbf{x}^*\|_2^2 \,|\mathcal{F}_k \right] \leq \mathbb{E}\left[r_1|\mathcal{F}_k\right] + \mathbb{E}\left[r_2|\mathcal{F}_k\right] + \alpha^2\|\boldsymbol{\pi}_r\|_2^2\overline{\boldsymbol{\pi}_c}\mathbb{E}\left[\|\mathbf{y}_k - B_\infty\mathbf{y}_k\|_{\boldsymbol{\pi}_c}^2 \,|\mathcal{F}_k \right] \tag{23}$$

Bounding $\mathbb{E}\left[r_1|\mathcal{F}_k\right]$: We first derive an upper bound on r_1. To simplify the notation, we denote $\widetilde{\alpha} \triangleq \alpha n\boldsymbol{\pi}_r^\top\boldsymbol{\pi}_c$. If $0 < \widetilde{\alpha} < \frac{1}{l}$, we have the following:

$$r_1 = \|\widehat{\mathbf{x}}_k - \mathbf{x}^* - \widetilde{\alpha}\nabla F(\widehat{\mathbf{x}}_k) + \widetilde{\alpha}\nabla F(\widehat{\mathbf{x}}_k) - \widetilde{\alpha}\overline{\mathbf{y}}_k\|_2^2$$

$$= \|\widehat{\mathbf{x}}_k - \mathbf{x}^* - \widetilde{\alpha}\nabla F(\widehat{\mathbf{x}}_k)\|_2^2 + \widetilde{\alpha}^2 \|\nabla F(\widehat{\mathbf{x}}_k) - \overline{\mathbf{y}}_k\|_2^2 + 2\widetilde{\alpha}\left\langle \widehat{\mathbf{x}}_k - \mathbf{x}^* - \widetilde{\alpha}\nabla F(\widehat{\mathbf{x}}_k), \nabla F(\widehat{\mathbf{x}}_k) - \overline{\mathbf{y}}_k \right\rangle$$

$$\leq (1 - \mu\widetilde{\alpha})^2 \|\widehat{\mathbf{x}}_k - \mathbf{x}^*\|_2^2 + \widetilde{\alpha}^2 \|\nabla F(\widehat{\mathbf{x}}_k) - \overline{\mathbf{y}}_k\|_2^2 + 2\widetilde{\alpha}\left\langle \widehat{\mathbf{x}}_k - \mathbf{x}^* - \widetilde{\alpha}\nabla F(\widehat{\mathbf{x}}_k), \nabla F(\widehat{\mathbf{x}}_k) - \overline{\mathbf{y}}_k \right\rangle,$$

where in the last inequality above we used Lemma 4. Then we take the conditional expectation given $\mathcal{F}_k$ on bothsides to obtain:

$$\mathbb{E}\left[r_1|\mathcal{F}_k\right]$$

$$= (1 - \mu\widetilde{\alpha})^2 \|\widehat{\mathbf{x}}_k - \mathbf{x}^*\|_2^2 + \widetilde{\alpha}^2\mathbb{E}\left[\|\nabla F(\widehat{\mathbf{x}}_k) - \overline{\mathbf{y}}_k\|_2^2 \,|\mathcal{F}_k \right] + 2\widetilde{\alpha}\left\langle \widehat{\mathbf{x}}_k - \mathbf{x}^* - \widetilde{\alpha}\nabla F(\widehat{\mathbf{x}}_k), \nabla F(\widehat{\mathbf{x}}_k) - \mathbf{h}\left(\mathbf{x}_k\right) \right\rangle$$

$$\leq (1 - \mu\widetilde{\alpha})^2 \|\widehat{\mathbf{x}}_k - \mathbf{x}^*\|_2^2 + \widetilde{\alpha}^2\mathbb{E}\left[\|\nabla F(\widehat{\mathbf{x}}_k) - \overline{\mathbf{y}}_k\|_2^2 \,|\mathcal{F}_k \right] + 2\widetilde{\alpha}\left(1 - \mu\widetilde{\alpha}\right)\|\widehat{\mathbf{x}}_k - \mathbf{x}^*\|_2 \|\nabla F(\widehat{\mathbf{x}}_k) - \mathbf{h}\left(\mathbf{x}_k\right)\|_2$$

$$\leq (1 - \mu\widetilde{\alpha})^2 \|\widehat{\mathbf{x}}_k - \mathbf{x}^*\|_2^2 + \widetilde{\alpha}^2\mathbb{E}\left[\|\nabla F(\widehat{\mathbf{x}}_k) - \overline{\mathbf{y}}_k\|_2^2 \,|\mathcal{F}_k \right]$$

$$\qquad + \widetilde{\alpha}\left(\mu\left(1 - \mu\widetilde{\alpha}\right)^2 \|\widehat{\mathbf{x}}_k - \mathbf{x}^*\|_2^2 + \frac{1}{\mu} \|\nabla F(\widehat{\mathbf{x}}_k) - \mathbf{h}\left(\mathbf{x}_k\right)\|_2^2 \right)$$

$$\leq (1 - \mu\widetilde{\alpha}) \|\widehat{\mathbf{x}}_k - \mathbf{x}^*\|_2^2 + \widetilde{\alpha}^2\mathbb{E}\left[\|\nabla F(\widehat{\mathbf{x}}_k) - \overline{\mathbf{y}}_k\|_2^2 \,|\mathcal{F}_k \right] + \frac{\widetilde{\alpha}l^2}{\mu n\underline{\boldsymbol{\pi}_r}} \|\mathbf{x}_k - \mathbf{1}_n\widehat{\mathbf{x}}_k\|_{\boldsymbol{\pi}_r}^2, \tag{24}$$

where in the second last inequality we used Lemma 4 and Young's inequality and in the last inequality we used Lemma 3 and the fact that $(1 - \mu\widetilde{\alpha})^2 (1 + \mu\widetilde{\alpha}) < (1 - \mu\widetilde{\alpha})$. In order to finish bounding r_1, we next bound $\mathbb{E}\left[\|\nabla F(\widehat{\mathbf{x}}_k) - \overline{\mathbf{y}}_k\|_2^2 \,|\mathcal{F}_k \right]$:

$$\mathbb{E}\left[\|\nabla F(\widehat{\mathbf{x}}_k) - \overline{\mathbf{y}}_k\|_2^2 \,|\mathcal{F}_k \right]$$

$$= \mathbb{E}\left[\|\nabla F(\widehat{\mathbf{x}}_k) - \mathbf{h}(\mathbf{x}_k) + \mathbf{h}(\mathbf{x}_k) - \overline{\mathbf{y}}_k\|_2^2 \,|\mathcal{F}_k \right]$$

$$= \|\nabla F(\widehat{\mathbf{x}}_k) - \mathbf{h}(\mathbf{x}_k)\|_2^2 + \mathbb{E}\left[\|\mathbf{h}(\mathbf{x}_k) - \overline{\mathbf{y}}_k\|_2^2 \,|\mathcal{F}_k \right]$$

Using Lemma 2 (3) and Lemma 3 leads to:

$$\mathbb{E}\left[\left\|\nabla F(\widehat{\mathbf{x}}_k) - \overline{\mathbf{y}}_k\right\|_2^2 \big| \mathcal{F}_k\right] \leq \frac{l^2}{n\underline{\boldsymbol{\pi}_r}}\left\|\mathbf{x}_k - \mathbf{1}_n\widehat{\mathbf{x}}_k\right\|_{\boldsymbol{\pi}_r}^2 + \frac{\sigma^2}{n} \tag{25}$$

Plugging in (25) to (24), we obtain an upper bound on $\mathbb{E}\left[r_1|\mathcal{F}_k\right]$ as follows:

$$\mathbb{E}\left[r_1|\mathcal{F}_k\right] \leq (1-\mu\widetilde{\alpha})\left\|\widehat{\mathbf{x}}_k - \mathbf{x}^*\right\|_2^2 + \frac{\widetilde{\alpha}l^2}{n\underline{\boldsymbol{\pi}_r}}\left(\frac{1}{\mu} + \widetilde{\alpha}\right)\left\|\mathbf{x}_k - \mathbf{1}_n\widehat{\mathbf{x}}_k\right\|_{\boldsymbol{\pi}_r}^2 + \frac{2\widetilde{\alpha}^2\sigma^2}{n}$$

$$\leq (1-\mu\widetilde{\alpha})\left\|\widehat{\mathbf{x}}_k - \mathbf{x}^*\right\|_2^2 + \frac{2\widetilde{\alpha}l^2}{\mu n\underline{\boldsymbol{\pi}_r}}\left\|\mathbf{x}_k - \mathbf{1}_n\widehat{\mathbf{x}}_k\right\|_{\boldsymbol{\pi}_r}^2 + \frac{\widetilde{\alpha}^2\sigma^2}{n} \tag{26}$$

Bounding $\mathbb{E}\left[r_2|\mathcal{F}_k\right]$: Recall that $r_2 = -2\alpha\langle\widehat{\mathbf{x}}_k - \mathbf{x}^* - \widetilde{\alpha}\overline{\mathbf{y}}_k, \boldsymbol{\pi}_r^\top(\mathbf{y}_k - B_\infty\mathbf{y}_k)\rangle$.

$$r_2 \leq 2\alpha\left\|\widehat{\mathbf{x}}_k - \mathbf{x}^* - \widetilde{\alpha}\overline{\mathbf{y}}_k\right\|_2\left\|\boldsymbol{\pi}_r^\top(\mathbf{y}_k - B_\infty\mathbf{y}_k)\right\|_2$$

$$\leq \alpha\left(\frac{\mu n\boldsymbol{\pi}_r^\top\boldsymbol{\pi}_c}{2}r_1 + \frac{2\left\|\boldsymbol{\pi}_r\right\|_2^2\overline{\boldsymbol{\pi}_c}}{\mu n\boldsymbol{\pi}_r^\top\boldsymbol{\pi}_c}\left\|\mathbf{y}_k - B_\infty\mathbf{y}_k\right\|_{\boldsymbol{\pi}_c}^2\right)$$

Taking the conditional expectation on bothsides given $\mathcal{F}_k$ to get:

$$\mathbb{E}\left[r_2|\mathcal{F}_k\right] \leq \frac{\mu\widetilde{\alpha}}{2}\mathbb{E}\left[r_1|\mathcal{F}_k\right] + \frac{2\alpha\left\|\boldsymbol{\pi}_r\right\|_2^2\overline{\boldsymbol{\pi}_c}}{\mu n\boldsymbol{\pi}_r^\top\boldsymbol{\pi}_c}\mathbb{E}\left[\left\|\mathbf{y}_k - B_\infty\mathbf{y}_k\right\|_{\boldsymbol{\pi}_c}^2\big|\mathcal{F}_k\right] \tag{27}$$

Bounding $\mathbb{E}\left[r_1|\mathcal{F}_k\right] + \mathbb{E}\left[r_2|\mathcal{F}_k\right]$: Putting (26) and (27) together, we have:

$$\mathbb{E}\left[r_1|\mathcal{F}_k\right] + \mathbb{E}\left[r_2|\mathcal{F}_k\right]$$

$$\leq \left(1 + \frac{\mu\widetilde{\alpha}}{2}\right)\frac{2\widetilde{\alpha}l^2}{\mu n\underline{\boldsymbol{\pi}_r}}\left\|\mathbf{x}_k - \mathbf{1}_n\widehat{\mathbf{x}}_k\right\|_{\boldsymbol{\pi}_r}^2 + \left(1 + \frac{\mu\widetilde{\alpha}}{2}\right)(1 - \mu\widetilde{\alpha})\left\|\widehat{\mathbf{x}}_k - \mathbf{x}^*\right\|_2^2 + \left(1 + \frac{\mu\widetilde{\alpha}}{2}\right)\frac{\widetilde{\alpha}^2\sigma^2}{n}$$

$$+ \frac{2\alpha\left\|\boldsymbol{\pi}_r\right\|_2^2\overline{\boldsymbol{\pi}_c}}{\mu n\boldsymbol{\pi}_r^\top\boldsymbol{\pi}_c}\left\|\mathbf{y}_k - B_\infty\mathbf{y}_k\right\|_{\boldsymbol{\pi}_c}^2$$

$$\leq \frac{3\widetilde{\alpha}l^2}{\mu n\underline{\boldsymbol{\pi}_r}}\left\|\mathbf{x}_k - \mathbf{1}_n\widehat{\mathbf{x}}_k\right\|_{\boldsymbol{\pi}_r}^2 + \left(1 - \frac{\mu\widetilde{\alpha}}{2}\right)\left\|\widehat{\mathbf{x}}_k - \mathbf{x}^*\right\|_2^2 + \frac{2\alpha\left\|\boldsymbol{\pi}_r\right\|_2^2\overline{\boldsymbol{\pi}_c}}{\mu n\boldsymbol{\pi}_r^\top\boldsymbol{\pi}_c}\left\|\mathbf{y}_k - B_\infty\mathbf{y}_k\right\|_{\boldsymbol{\pi}_c}^2 + \frac{3\widetilde{\alpha}^2\sigma^2}{2n}, \tag{28}$$

where in the last inequality we use the fact that $1 + \frac{\mu\widetilde{\alpha}}{2} \leq \frac{3}{2}$ and $\left(1 + \frac{\mu\widetilde{\alpha}}{2}\right)(1 - \mu\widetilde{\alpha}) \leq 1 - \frac{\mu\widetilde{\alpha}}{2}$. Plugging (28) in (23) and replacing $\widetilde{\alpha}$ by $\alpha n\boldsymbol{\pi}_r^\top\boldsymbol{\pi}_c$ finishes the proof. $\blacksquare$

PROOF OF LEMMA 9

Proof: To simplify notation, we define:

$$\nabla_k \triangleq \nabla\mathbf{f}(\mathbf{x}_k), \qquad \nabla_k^i \triangleq \nabla f_i(\mathbf{x}_k^i),$$

$$\mathbf{g}_k \triangleq \mathbf{g}(\mathbf{x}_k, \boldsymbol{\xi}_k), \qquad \mathbf{g}_k^i \triangleq \mathbf{g}_i(\mathbf{x}_k^i, \xi_k^i).$$

Starting with (3b) and using Lemma 1, we obtain

$$\left\|\mathbf{y}_{k+1} - B_\infty\mathbf{y}_{k+1}\right\|_{\boldsymbol{\pi}_c}^2$$

$$= \left\|B\mathbf{y}_k - B_\infty\mathbf{y}_k\right\|_{\boldsymbol{\pi}_c}^2 + \left\|(I_n - B_\infty)(\mathbf{g}_{k+1} - \mathbf{g}_k)\right\|_{\boldsymbol{\pi}_c}^2 + 2\left\langle B\mathbf{y}_k - B_\infty\mathbf{y}_k, (I_n - B_\infty)(\mathbf{g}_{k+1} - \mathbf{g}_k)\right\rangle_{\boldsymbol{\pi}_c}$$

$$\leq \sigma_B^2\left\|\mathbf{y}_k - B_\infty\mathbf{y}_k\right\|_{\boldsymbol{\pi}_c}^2 + \left\|\mathbf{g}_{k+1} - \mathbf{g}_k\right\|_{\boldsymbol{\pi}_c}^2 + 2\left\langle B\mathbf{y}_k - B_\infty\mathbf{y}_k, \mathbf{g}_{k+1} - \mathbf{g}_k\right\rangle_{\boldsymbol{\pi}_c}, \tag{29}$$

where the last inequality uses $\left\|\!\left\| I_n - B_\infty \right\|\!\right\|_{\boldsymbol{\pi}_c} = 1$ and that

$$\left\langle B\mathbf{y}_k - B_\infty \mathbf{y}_k, B_\infty \left(\mathbf{g}_{k+1} - \mathbf{g}_k\right) \right\rangle_{\boldsymbol{\pi}_c} = \left\langle B\mathbf{y}_k - B_\infty \mathbf{y}_k, \mathbf{1}_n \mathbf{1}_n^\top \left(\mathbf{g}_{k+1} - \mathbf{g}_k\right) \right\rangle = 0.$$

We take the conditional expectation given $\mathcal{F}_k$ on both sides of (29) to get:

$$\mathbb{E}\left[\left\|\mathbf{y}_{k+1} - B_\infty \mathbf{y}_{k+1}\right\|_{\boldsymbol{\pi}_c}^2 \big| \mathcal{F}_k\right]$$

$$\leq \sigma_B^2 \mathbb{E}\left[\left\|\mathbf{y}_k - B_\infty \mathbf{y}_k\right\|_{\boldsymbol{\pi}_c}^2 \big| \mathcal{F}_k\right] + \mathbb{E}\left[\left\|\mathbf{g}_{k+1} - \mathbf{g}_k\right\|_{\boldsymbol{\pi}_c}^2 \big| \mathcal{F}_k\right] + 2\mathbb{E}\left[\mathbb{E}\left[\left\langle B\mathbf{y}_k - B_\infty \mathbf{y}_k, \mathbf{g}_{k+1} - \mathbf{g}_k\right\rangle_{\boldsymbol{\pi}_c} \big| \mathcal{F}_{k+1}\right] \big| \mathcal{F}_k\right]$$

$$= \sigma_B^2 \mathbb{E}\left[\left\|\mathbf{y}_k - B_\infty \mathbf{y}_k\right\|_{\boldsymbol{\pi}_c}^2 \big| \mathcal{F}_k\right] + \mathbb{E}\left[\left\|\mathbf{g}_{k+1} - \mathbf{g}_k\right\|_{\boldsymbol{\pi}_c}^2 \big| \mathcal{F}_k\right] + 2\mathbb{E}\left[\left\langle B\mathbf{y}_k - B_\infty \mathbf{y}_k, \nabla_k - \mathbf{g}_k\right\rangle_{\boldsymbol{\pi}_c} \big| \mathcal{F}_k\right]$$

$$+ 2\mathbb{E}\left[\left\langle B\mathbf{y}_k - B_\infty \mathbf{y}_k, \nabla_{k+1} - \nabla_k\right\rangle_{\boldsymbol{\pi}_c} \big| \mathcal{F}_k\right],$$

$$\triangleq \sigma_B^2 \mathbb{E}\left[\left\|\mathbf{y}_k - B_\infty \mathbf{y}_k\right\|_{\boldsymbol{\pi}_c}^2 \big| \mathcal{F}_k\right] + s_1 + 2s_2 + s_3. \tag{30}$$

We now bound the last three terms in the following. We start with s_1.

Bounding s_1:

$$s_1 = \mathbb{E}\left[\left\|\mathbf{g}_{k+1} - \mathbf{g}_k - (\nabla_{k+1} - \nabla_k) + \nabla_{k+1} - \nabla_k\right\|_{\boldsymbol{\pi}_c}^2 \big| \mathcal{F}_k\right]$$

$$= \mathbb{E}\left[\left\|\nabla_{k+1} - \nabla_k\right\|_{\boldsymbol{\pi}_c}^2 \big| \mathcal{F}_k\right]$$

$$+ \mathbb{E}\left[\left\|\mathbf{g}_{k+1} - \mathbf{g}_k - (\nabla_{k+1} - \nabla_k)\right\|_{\boldsymbol{\pi}_c}^2 \big| \mathcal{F}_k\right] + 2\mathbb{E}\left[\left\langle \nabla_{k+1} - \nabla_k, \mathbf{g}_{k+1} - \mathbf{g}_k - (\nabla_{k+1} - \nabla_k)\right\rangle_{\boldsymbol{\pi}_c} \big| \mathcal{F}_k\right]$$

$$\leq \mathbb{E}\left[\left\|\nabla_{k+1} - \nabla_k\right\|_{\boldsymbol{\pi}_c}^2 \big| \mathcal{F}_k\right] + 2\mathbb{E}\left[\left\langle \nabla_{k+1}, \nabla_k - \mathbf{g}_k\right\rangle_{\boldsymbol{\pi}_c} \big| \mathcal{F}_k\right] + \frac{2n\sigma^2}{\underline{\boldsymbol{\pi}_c}}. \tag{31}$$

We now bound the first term in the above inequality. Using (3a), Lemma 1, we have

$$\left\|\nabla_{k+1} - \nabla_k\right\|_{\boldsymbol{\pi}_c}^2$$

$$\leq \tfrac{l^2}{\underline{\boldsymbol{\pi}_c}} \left\|\mathbf{x}_{k+1} - \mathbf{x}_k\right\|_2^2$$

$$= \tfrac{l^2}{\underline{\boldsymbol{\pi}_c}} \left\|A\mathbf{x}_k - \alpha\mathbf{y}_k - \mathbf{x}_k\right\|_2^2$$

$$= \tfrac{l^2}{\underline{\boldsymbol{\pi}_c}} \left\|(A - I_n)(\mathbf{x}_k - A_\infty \mathbf{x}_k) - \alpha\mathbf{y}_k\right\|_2^2$$

$$\leq \tfrac{8l^2}{\underline{\boldsymbol{\pi}_c}\,\underline{\boldsymbol{\pi}_r}} \left\|\mathbf{x}_k - A_\infty \mathbf{x}_k\right\|_{\boldsymbol{\pi}_r}^2 + \tfrac{2\alpha^2 l^2}{\underline{\boldsymbol{\pi}_c}} \left\|\mathbf{y}_k\right\|_2^2, \tag{32}$$

where in the last inequality we used the basic inequality $\left\|\mathbf{x} + \mathbf{y}\right\|_2^2 \leq 2\left\|\mathbf{x}\right\|_2^2 + 2\left\|\mathbf{y}\right\|_2^2, \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^p$ and that $\left\|\!\left\| A - I_n \right\|\!\right\|_{\boldsymbol{\pi}_r} \leq 2$. In order to bound the second term in (31), we first note, from (2a), that

$$\nabla_{k+1}^i = \nabla f_i\left(\sum_{j=1}^n a_{ij}\mathbf{x}_k^j - \alpha\left(\sum_{j=1}^n b_{ij}\mathbf{y}_{k-1}^j + \mathbf{g}_k^i - \mathbf{g}_{k-1}^i\right)\right),$$

and we also define

$$\widetilde{\nabla}_{k+1}^i \triangleq \nabla f_i\left(\sum_{j=1}^n a_{ij}\mathbf{x}_k^j - \alpha\left(\sum_{j=1}^n b_{ij}\mathbf{y}_{k-1}^j + \nabla_k^i - \mathbf{g}_{k-1}^i\right)\right).$$

Therefore, $\left\|\nabla_{k+1}^i - \widetilde{\nabla}_{k+1}^i\right\|_2 \le \alpha l \left\|\nabla_k^i - \mathbf{g}_k^i\right\|_2$. We then proceed to bound the second term in (31) as follows:

$$
\mathbb{E}\left[\left\langle \nabla_{k+1}, \nabla_k - \mathbf{g}_k \right\rangle_{\boldsymbol{\pi}_c} \big| \mathcal{F}_k\right]
$$

$$
\le \frac{1}{\underline{\boldsymbol{\pi}_c}} \sum_{i=1}^n \mathbb{E}\left[\left\langle \nabla_{k+1}^i - \widetilde{\nabla}_k^i + \widetilde{\nabla}_k^i, \nabla_k^i - \mathbf{g}_k^i \right\rangle \big| \mathcal{F}_k\right]
$$

$$
\le \frac{1}{\underline{\boldsymbol{\pi}_c}} \sum_{i=1}^n \mathbb{E}\left[\left\|\nabla_{k+1}^i - \widetilde{\nabla}_k^i\right\|_2 \left\|\nabla_k^i - \mathbf{g}_k^i\right\|_2 \big| \mathcal{F}_k\right]
$$

$$
\le \frac{\alpha l}{\underline{\boldsymbol{\pi}_c}} \sum_{i=1}^n \mathbb{E}\left[\left\|\nabla_k^i - \mathbf{g}_k^i\right\|_2^2 \big| \mathcal{F}_k\right]
$$

$$
\le \frac{\alpha n l \sigma^2}{\underline{\boldsymbol{\pi}_c}}. \tag{33}
$$

Plugging (32) and (33) to (31), we obtain an upper bound on s_1:

$$
s_1 \le \frac{8l^2}{\underline{\boldsymbol{\pi}_c}\,\underline{\boldsymbol{\pi}_r}} \left\|\mathbf{x}_k - A_\infty \mathbf{x}_k\right\|_{\boldsymbol{\pi}_r}^2 + \frac{2\alpha^2 l^2}{\underline{\boldsymbol{\pi}_c}} \mathbb{E}\left[\left\|\mathbf{y}_k\right\|_2^2 \big| \mathcal{F}_k\right] + \frac{2\alpha l n \sigma^2}{\underline{\boldsymbol{\pi}_c}} + \frac{2n\sigma^2}{\underline{\boldsymbol{\pi}_c}}. \tag{34}
$$

Bounding s_2: We first split s_2 into the sum of two terms as follows:

$$
s_2 = \mathbb{E}\left[\left\langle B\mathbf{y}_k, \nabla_k - \mathbf{g}_k \right\rangle_{\boldsymbol{\pi}_c} \big| \mathcal{F}_k\right] - \mathbb{E}\left[\left\langle B_\infty \mathbf{y}_k, \nabla_k - \mathbf{g}_k \right\rangle_{\boldsymbol{\pi}_c} \big| \mathcal{F}_k\right]. \tag{35}
$$

For the first term in (35), using (3b), we have

$$
\mathbb{E}\left[\left\langle B\mathbf{y}_k, \nabla_k - \mathbf{g}_k \right\rangle_{\boldsymbol{\pi}_c} \big| \mathcal{F}_k\right]
$$

$$
= \mathbb{E}\left[\left\langle B^2 \mathbf{y}_{k-1} + B(\mathbf{g}_k - \mathbf{g}_{k-1}), \nabla_k - \mathbf{g}_k \right\rangle_{\boldsymbol{\pi}_c} \big| \mathcal{F}_k\right]
$$

$$
= \mathbb{E}\left[\left\langle B\mathbf{g}_k, \nabla_k - \mathbf{g}_k \right\rangle_{\boldsymbol{\pi}_c} \big| \mathcal{F}_k\right]
$$

$$
= \sum_{i=1}^n \frac{1}{[\boldsymbol{\pi}_c]_i} \mathbb{E}\left[\left\langle \sum_{j=1}^n b_{ij} \mathbf{g}_k^j, \nabla_k^i - \mathbf{g}_k^i \right\rangle \big| \mathcal{F}_k\right]
$$

$$
= \sum_{i=1}^n \frac{b_{ii}}{[\boldsymbol{\pi}_c]_i} \mathbb{E}\left[\left\langle \mathbf{g}_k^i, \nabla_k^i - \mathbf{g}_k^i \right\rangle \big| \mathcal{F}_k\right] \le 0.
$$

For the second term in (35), we have

$$
-\mathbb{E}\left[\left\langle B_\infty \mathbf{y}_k, \nabla_k - \mathbf{g}_k \right\rangle_{\boldsymbol{\pi}_c} \big| \mathcal{F}_k\right]
$$

$$
= -\mathbb{E}\left[\left\langle \mathbf{1}_n \mathbf{1}_n^\top \mathbf{g}_k, \nabla_k - \mathbf{g}_k \right\rangle \big| \mathcal{F}_k\right]
$$

$$
= \sum_{i=1}^n \mathbb{E}\left[\left\langle \sum_{j=1}^n -\mathbf{g}_k^j, \nabla_k^i - \mathbf{g}_k^i \right\rangle \big| \mathcal{F}_k\right]
$$

$$
= \sum_{i=1}^n \mathbb{E}\left[\left\langle \nabla_k^i - \mathbf{g}_k^i, \nabla_k^i - \mathbf{g}_k^i \right\rangle \big| \mathcal{F}_k\right] \le n\sigma^2
$$

Hence, we have an upper bound on s_2 as follows:

$$
s_2 \le n\sigma^2. \tag{36}
$$

Bounding s_3: Using the upper bound on $\|\nabla_{k+1} - \nabla_k\|_{\boldsymbol{\pi}_c}^2$ in (32), we proceed towards an upper bound on s_3.

$$2\langle B\mathbf{y}_k - B_\infty\mathbf{y}_k, \nabla_{k+1} - \nabla_k\rangle_{\boldsymbol{\pi}_c}$$

$$\leq 2\|B\mathbf{y}_k - B_\infty\mathbf{y}_k\|_{\boldsymbol{\pi}_c}\|\nabla_{k+1} - \nabla_k\|_{\boldsymbol{\pi}_c}$$

$$\leq \frac{1 - \sigma_B^2}{2\sigma_B^2}\|B\mathbf{y}_k - B_\infty\mathbf{y}_k\|_{\boldsymbol{\pi}_c}^2 + \frac{2\sigma_B^2}{1 - \sigma_B^2}\|\nabla_{k+1} - \nabla_k\|_{\boldsymbol{\pi}_c}^2$$

$$\leq \frac{1 - \sigma_B^2}{2}\|\mathbf{y}_k - B_\infty\mathbf{y}_k\|_{\boldsymbol{\pi}_c}^2 + \frac{16\sigma_B^2 l^2}{\underline{\boldsymbol{\pi}_c}\,\underline{\boldsymbol{\pi}_r}(1 - \sigma_B^2)}\|\mathbf{x}_k - A_\infty\mathbf{x}_k\|_{\boldsymbol{\pi}_r}^2 + \frac{4\sigma_B^2 l^2\alpha^2}{\underline{\boldsymbol{\pi}_c}(1 - \sigma_B^2)}\|\mathbf{y}_k\|_2^2.$$

Taking the conditional expectation given $\mathcal{F}_k$ on bothsides of the inequality above, we obtain an upper bound on s_3 as follows:

$$s_3 \leq \frac{16\sigma_B^2 l^2}{\underline{\boldsymbol{\pi}_c}\,\underline{\boldsymbol{\pi}_r}(1 - \sigma_B^2)}\|\mathbf{x}_k - A_\infty\mathbf{x}_k\|_{\boldsymbol{\pi}_r}^2 + \frac{1 - \sigma_B^2}{2}\mathbb{E}\left[\|\mathbf{y}_k - B_\infty\mathbf{y}_k\|_{\boldsymbol{\pi}_c}^2\,\big|\mathcal{F}_k\right] + \frac{4\sigma_B^2 l^2\alpha^2}{\underline{\boldsymbol{\pi}_c}(1 - \sigma_B^2)}\mathbb{E}\left[\|\mathbf{y}_k\|_2^2\,\big|\mathcal{F}_k\right] \quad (37)$$

■

Plugging the bounds on s_1, s_2, s_3 in (34), (36) and (37) into (30) completes the proof.