VARIANCE REDUCED STOCHASTIC OPTIMIZATION OVER DIRECTED GRAPHS WITH ROW AND COLUMN STOCHASTIC WEIGHTS

Muhammad I. Qureshi[†]*, Ran Xin*[‡]*, Soummya Kar*[‡]*, and Usman A. Khan*[†]

[†]Tufts University, MA, USA, [‡]Carnegie Mellon University, PA, USA

2023 57th Asilomar Conference on Signals, Systems, and Computers | 979-8-3503-2574-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/IEEECONF59524.2023.10476854

ABSTRACT

This paper proposes **AB-SAGA**, a first-order distributed stochastic optimization method to minimize a finite sum of smooth and strongly convex functions distributed over an arbitrary directed graph. **AB-SAGA** removes the uncertainty caused by the stochastic gradients using node-level variance reduction and subsequently employs network-level gradient tracking to address the data dissimilarity across the nodes. Unlike existing methods that use the nonlinear push-sum correction to cancel the imbalance caused by the directed communication, the consensus updates in **AB-SAGA** are linear and use both row and column stochastic weights. We show that for a constant stepsize, **AB-SAGA** converges linearly to the global optimal. We quantify the directed nature of the underlying graph using an explicit directivity constant and characterize the regimes in which **AB-SAGA** achieves a linear speed-up over its centralized counterpart. Numerical experiments illustrate the convergence of **AB-SAGA** for strongly convex and non-convex problems.

***Index Terms*—** Stochastic optimization, variance reduction, first-order methods, directed graphs.

1. INTRODUCTION

Stochastic optimization is relevant in many signal processing, machine learning, and control applications [1, 2]. In large-scale problems, data is usually geographically distributed making centralized methods practically infeasible. Distributed solutions are thus preferable where individual nodes perform local updates using data fusion among the nearby nodes [3–5]. The problem of interest can be written as

$$\mathbf{P}: \quad \min_{\mathbf{x}\in\mathbb{R}^p} F(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}), \;\; f_i(\mathbf{x}) := \frac{1}{m_i}\sum_{j=1}^{m_i} f_{i,j}(\mathbf{x}),$$

where each local cost function $f_i : \mathbb{R}^p \to \mathbb{R}$ is private to node i and is further decomposed into m_i component cost functions $\{f_{i,j} : \mathbb{R}^p \to \mathbb{R}\}_{j=1}^{m_i}$. When the underlying problem is smooth and strongly convex, the goal is to find the unique minimizer $\mathbf{x}^*$ of the global cost F, given that the nodes communicate over a strongly connected directed graph.

Distributed first-order stochastic methods for problem $\mathbf{P}$ are well studied in the literature. Early work includes [6, 7] that is applicable to undirected graphs. **SGP** (stochastic gradient push [8]) extends **DSGD** (distributed stochastic gradient descent [6]) to directed graphs using push-sum consensus [9]. Both **DSGD** and **SGP** suffer from a steady-state error caused by the difference in global and local cost functions, i.e., $\|\nabla F(\mathbf{x}^*) - \nabla f_i(\mathbf{x}^*)\|$, and the variance introduced by the stochastic gradients. Over arbitrary directed graphs, **S-ADDOPT** [10] compensates for the heterogeneity of local cost functions using gradient tracking [11–13]. However, the steady-state error remains in effect due to the variance. A recent work **Push-SAGA** [14] benefits from a variance reduction technique [15] to eliminate this error. Both **S-ADDOPT** and **Push-SAGA** use nonlinear push-sum corrections to ensure agreement by dividing by the estimates of the right Perron eigenvector of the underlying column stochastic weight matrix which slows down the convergence and adds to the computational cost. Such correction is not required when the weights are doubly stochastic as is the case over undirected graphs; see [16–20] for related work.

In this paper, we present **AB-SAGA**, a first-order distributed stochastic optimization method that is applicable to arbitrary directed graphs. Similar to [14, 18], **AB-SAGA** eliminates the uncertainty caused by the stochastic gradients using variance reduction and addresses the global versus local cost gaps using gradient tracking. Unlike **Push-SAGA** [14], however, **AB-SAGA** uses both row and column stochastic weights for consensus, thus eliminating the need to estimate the Perron eigenvector required in push-sum methods. The main contributions of this paper are summarized next: (i) We analytically establish the linear convergence of **AB-SAGA** to the global optimizer $\mathbf{x}^*$ for smooth and strongly convex problems; (ii) We quantify the performance of **AB-SAGA** over directed graphs and encapsulate the directed nature of the communication in a directivity constant $\psi \geq 1$, which is 1 for undirected graphs; (iii) We provide explicit expressions for the gradient computation and communication complexities, and show that **AB-SAGA** achieves linear speedup over its centralized counterpart.

We now describe the rest of the paper. Section 2 motivates the algorithm development and formally describes **AB-SAGA**. Section 3 describes the assumptions and the main results. Section 4 provides the detailed convergence analysis. Section 5 presents the numerical experiments on strongly convex and non-convex problems. Section 6 concludes the paper.

Basic Notation: We use upper case letters to represent matrices and lower case bold letters for vectors. We define I_n as $n \times n$ identity matrix and $\mathbf{1}_n$ as a column vector of n ones. From Perron Frobenius theorem [21], for a primitive row stochastic matrix $\underline{A} \in \mathbb{R}^{n\times n}$ (column stochastic matrix $\underline{B} \in \mathbb{R}^{n\times n}$), we define $\underline{A}^\infty := \lim_{k\to\infty} \underline{A}^k = \boldsymbol{\pi}_r^\top \mathbf{1}_n$ ($\underline{B}^\infty := \lim_{k\to\infty} \underline{B}^k = \mathbf{1}_n^\top \boldsymbol{\pi}_c$), where $\boldsymbol{\pi}_r$ is the left eigenvector of $\underline{A}$ ($\boldsymbol{\pi}_c$ is the right eigenvector of $\underline{B}$), corresponding to the unique eigenvalue 1. We further denote the largest element of a vector $\boldsymbol{\pi}_r$ as $\overline{\pi}_r$ and the smallest element as $\underline{\pi}_r$, and define the ratios $h_r := \overline{\pi}_r/\underline{\pi}_r$ and $h_c := \overline{\pi}_c/\underline{\pi}_c$. We denote the spectral radius of matrix $\underline{A}$ as $\rho(\underline{A})$, while $\|\cdot\|_2$ denotes the vector two-norm and $|||\cdot|||_2$ denotes the matrix norm induced by this vector norm. Since $\rho(\underline{A} - \underline{A}^\infty) < 1$ and $\rho(\underline{B} - \underline{B}^\infty) < 1$, it can be shown that there exist matrix norms $|||\cdot|||_{\boldsymbol{\pi}_r}$ and $|||\cdot|||_{\boldsymbol{\pi}_c}$, formally defined in [22], such that $\sigma_A := |||\underline{A} - \underline{A}^\infty|||_{\boldsymbol{\pi}_r} < 1$ and $\sigma_B := |||\underline{B} - \underline{B}^\infty|||_{\boldsymbol{\pi}_c} < 1$.

2. ALGORITHM DEVELOPMENT

We motivate the proposed algorithm with the help of a recent work **GT-DSGD** [17, 23], which adds gradient tracking to the well-known **DSGD** [6]. The **GT-DSGD** algorithm can be described as follows. Let $\underline{W} = \{\underline{w}_{ir}\}$ be the network weight matrix such that $\underline{w}_{ir} \neq 0$, if and only if node i can receive information from node r. Let $\mathbf{x}_i^k, \mathbf{w}_i^k$, both in $\mathbb{R}^p$, be the state vectors at each node i and iteration k. Then, $\forall k \geq 0$, **GT-DSGD** at each node is given by

$$\mathbf{x}_i^{k+1} = \sum_{r=1}^n \underline{w}_{ir}\mathbf{x}_r^k - \alpha \mathbf{w}_i^k,$$

$$\mathbf{w}_i^{k+1} = \sum_{r=1}^n \underline{w}_{ir}\mathbf{w}_r^k + \nabla f_{i,s_i^{k+1}}\left(\mathbf{x}_i^{k+1}\right) - \nabla f_{i,s_i^k}\left(\mathbf{x}_i^k\right),$$

where s_i^k is an index drawn uniformly at random from the index set $\{1, \cdots, m_i\}$ and $\nabla f_{i,s_i^k}(\mathbf{x}_i^k)$ is the gradient of the s_i^k-th component cost function f_{i,s_i^k} (and not the full local gradient ∇f_i). The $\mathbf{w}_i^k$-update in **GT-DSGD** is based on dynamic average consensus [24] and essentially tracks the global gradient ∇F, asymptotically, see [11–13] for more details. The $\mathbf{x}_i^k$-update consequently implements a descent in the global gradient direction $\mathbf{w}_i^k$. Assuming that the variance of local stochastic gradients is bounded, i.e., $\mathbb{E}_{s_i^k}[\|\nabla f_{i,s_i^k}(\mathbf{x}_i^k) - \nabla f_i(\mathbf{x}_i^k)\|_2^2 \mid \mathbf{x}_i^k] \leq \sigma^2$, and the global cost is ℓ-smooth and μ-strongly convex, **GT-DSGD** converges linearly to the neighborhood of the optimal solution, i.e.,

$$\limsup_{k\to\infty} \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|\mathbf{x}_i^k - \mathbf{x}^*\|_2^2] = \mathcal{O}\left(\tfrac{\alpha}{n\mu}\,\sigma^2 + \tfrac{\alpha^2\kappa^2}{(1-\lambda)^3}\,\sigma^2\right),$$

for a sufficiently small constant stepsize α, where $(1-\lambda)$ is the spectral gap of $\underline{W}$ and κ is the condition number of F. We note that the steady-state error depends on the variance σ^2 of the stochastic gradients and **GT-DSGD**, in general, is not applicable to arbitrary directed graphs since it requires $\underline{W}$ to be doubly stochastic.

Algorithm 1 **AB-SAGA** at each node i

Require: $\mathbf{x}_i^0 \in \mathbb{R}^p$, $\mathbf{w}_i^0 = \mathbf{g}_i^0 = \nabla f_i(\mathbf{x}_i^0)$, $\mathbf{v}_{i,j}^1 = \mathbf{x}_i^0$, $\forall j \in \{1, \cdots, m_i\}, \alpha > 0$, $\{a_{ir}\}_{r=1}^n$, $\{b_{ir}\}_{r=1}^n$, Gradient table: $\{\nabla f_{i,j}(\mathbf{x}_i^0)\}_{j=1}^{m_i}$

1: **for** $k = 0, 1, 2, \ldots$ **do**
2: $\quad \mathbf{x}_i^{k+1} \leftarrow \sum_{r=1}^n a_{ir}(\mathbf{x}_r^k - \alpha \cdot \mathbf{w}_r^k)$
3: $\quad$ **Select** s_i^{k+1} uniformly at random from $\{1, \ldots, m_i\}$
4: $\quad \mathbf{g}_i^{k+1} \leftarrow \nabla f_{i,s_i^{k+1}}(\mathbf{x}_i^{k+1}) - \nabla f_{i,s_i^{k+1}}(\mathbf{v}_{i,s_i^{k+1}}^{k+1}) + \frac{1}{m_i}\sum_{j=1}^{m_i} \nabla f_{i,j}(\mathbf{v}_{i,j}^{k+1})$
5: $\quad$ **Replace** $\nabla f_{i,s_i^{k+1}}(\mathbf{v}_{i,s_i^{k+1}}^{k+1})$ by $\nabla f_{i,s_i^{k+1}}(\mathbf{x}_i^{k+1})$ in the gradient table
6: $\quad \mathbf{w}_i^{k+1} \leftarrow \sum_{r=1}^n b_{ir}(\mathbf{w}_r^k + \mathbf{g}_r^{k+1} - \mathbf{g}_r^k)$
7: $\quad$ **if** $j = s_i^{k+1}$, **then** $\mathbf{v}_{i,j}^{k+2} \leftarrow \mathbf{x}_i^{k+1}$, **else** $\mathbf{v}_{i,j}^{k+2} \leftarrow \mathbf{v}_{i,j}^{k+1}$
8: $\quad$ **end if**
9: **end for**

In this paper, we propose **AB-SAGA** that removes the steady state error in **GT-DSGD** with the help of a **SAGA**-based variance reduction technique [15]. The complete implementation details are provided in Algorithm 1. We note that for each $\mathbf{x}_i^k$ and $\mathbf{w}_i^k$ updates, **AB-SAGA** requires $c \in \mathbb{N}$ and $d \in \mathbb{N}$ communication rounds, respectively. For ease of notation, we write the ir-th element of $\underline{A}^c$ as $\{a_{ir}\}$ and $\underline{B}^d$ as $\{b_{ir}\}$, for some $c, d \in \mathbb{N}$, formally defined later. Each node i updates $\mathbf{x}_i^k$, which estimates the global minimum $\mathbf{x}^*$, and $\mathbf{w}_i^k$, which tracks the global gradient $\nabla F(\mathbf{x}_i^k)$ using **SAGA**-based local gradient update $\mathbf{g}_i^{k+1}$. We remark that each node requires additional storage $\mathcal{O}(pm_i)$ to maintain the gradient table $\{\nabla f_{i,j}(\mathbf{v}_{i,j}^k)\}_{j=1}^{m_i}$ as is standard in **SAGA**. This storage cost can be reduced to $\mathcal{O}(m_i)$ for certain problems [15]. Finally, we note that **AB-SAGA** is applicable to arbitrary directed graphs as it does not require the corresponding weight matrices, $\underline{A}$ and $\underline{B}$, to be doubly stochastic.

3. ASSUMPTIONS AND MAIN RESULTS

We first describe the assumptions below.

Assumption 1. *The network of nodes communicates over a strongly connected arbitrary directed graph.*

Assumption 2. *The global cost function F is μ-strongly convex and each component cost $f_{i,j}$ is ℓ-smooth.*

Assumption 1 ensures that the matrices $\underline{A} = \{\underline{a}_{ir}\}$ and $\underline{B} = \{\underline{b}_{ir}\}$ can be chosen to be irreducible, primitive, row and column stochastic, respectively. In particular, if each node i has the knowledge of its in-degree d_i^{in} and its out-degree d_i^{out}, then the weights can be locally chosen as $\underline{a}_{ir} = 1/d_i^{\text{in}}$ for each incoming neighbor r, and $\underline{b}_{ir} = 1/d_i^{\text{out}}$ for each outgoing neighbor r. Next, Assumption 2 ensures that the global cost F has a unique minimizer $\mathbf{x}^*$. We note that the local cost functions f_i's are not necessarily strongly convex, which is a relaxed condition than the one for **Push-SAGA**. Based on these assumptions, we now present our main results.

Theorem 1. *Consider problem* $\mathbf{P}$ *under Assumptions 1 and 2. For the stepsize* $\alpha \in (0, \overline{\alpha}]$, $\overline{\alpha} > 0$ *formally defined later,* **AB-SAGA** *linearly converges to* $\mathbf{x}^*$. *In particular, when* $\alpha = \overline{\alpha}$, **AB-SAGA** *achieves an* ϵ*-optimal solution in*

$$\Gamma = \mathcal{O}\left(\max\left\{\kappa\psi, \tfrac{\kappa^2 M}{m}, M\right\}\log\tfrac{1}{\epsilon}\right)$$

gradient computations, with $(c+d)$ *communication rounds per iteration, for all* $c = \lceil\, \overline{c}\,\rceil$ *and* $d = \lceil\, \overline{d}\,\rceil$ *such that*

$$\overline{c} := \frac{\log\left(\frac{90512nM\kappa}{m(1-\sigma_B^{2d})}\sqrt{\frac{h_r h_c}{\boldsymbol{\pi}_r^\top\boldsymbol{\pi}_c}}\right)}{\log\frac{1}{\sigma_A}}, \quad \overline{d} := \frac{\log\left(\frac{1265\kappa}{\boldsymbol{\pi}_r^\top\boldsymbol{\pi}_c}\sqrt{\frac{nMh_c}{m}}\right)}{\log\frac{1}{\sigma_B}},$$

where $M := \max_i m_i, m := \min_i m_i, \kappa := \ell/\mu$ *is the condition number, and* $\psi := \frac{\sqrt{h_r h_c}}{n(\boldsymbol{\pi}_r^\top\boldsymbol{\pi}_c)}$ *is the directivity constant.*

The formal proof of the Theorem 1 is provided in Section 4. The following remarks summarize its key attributes.

Remark 1. *We note that for well-connected networks, i.e., when* σ_A *and* σ_B *are small, we have that* $\overline{c} \leq 1$ *and* $\overline{d} \leq 1$. *Thus, we get* $c = 1$ *and* $d = 1$, *and* **AB-SAGA** *converges with a single round of communication per iteration.*

Remark 2. *Theorem 1 describes an explicit directivity constant* $\psi := \frac{\sqrt{h_r h_c}}{n(\boldsymbol{\pi}_r^\top\boldsymbol{\pi}_c)}$, *which is* 1 *for undirected networks. Thus* **AB-SAGA** *and its convergence proof are naturally applicable to undirected graphs.*

Remark 3. *When each node possess a large dataset such that* $M \approx m \gg \kappa^2\psi$, **AB-SAGA** *achieves an* ϵ*-optimal solution in* $\mathcal{O}\left(M\log\epsilon^{-1}\right)$ *gradient computations per node. We note that this complexity is* n *times better than the centralized complexity* $\mathcal{O}\left(nM\log\epsilon^{-1}\right)$ *of* **SAGA** *[15] that processes all data at a single location.*

4. CONVERGENCE OF AB-SAGA

In this section, we formalize the convergence analysis. It can be verified that **AB-SAGA** described in Algorithm 1 can be compactly written in a vector-matrix format as

$$\mathbf{x}^{k+1} = A^c(\mathbf{x}^k - \alpha\mathbf{w}^k), \tag{1a}$$

$$\mathbf{w}^{k+1} = B^d(\mathbf{w}^k + \mathbf{g}^{k+1} - \mathbf{g}^k); \tag{1b}$$

where $\mathbf{x}^k, \mathbf{w}^k$ and $\mathbf{g}^k$ are the global state vectors in $\mathbb{R}^{pn}$ concatenating the local state vectors $\mathbf{x}_i^k, \mathbf{w}_i^k$ and $\mathbf{g}_i^k$ in $\mathbb{R}^p$, respectively. Similarly, $A := \underline{A} \otimes I_p$ and $B := \underline{B} \otimes I_p$, in $\mathbb{R}^{pn\times pn}$, are the global weight matrices, whereas c and d denote the communication rounds per iterate. We next define four error terms to aid the convergence analysis of **AB-SAGA**:

(i) Network agreement error: $\mathbb{E}\|\mathbf{x}^k - A^\infty\mathbf{x}^k\|^2$;
(ii) Optimality gap: $\mathbb{E}\|\widehat{\mathbf{x}}^k - \mathbf{x}^*\|^2$;
(iii) Mean auxiliary gap: $\mathbb{E}[\mathbf{t}^k]$;
(iv) Gradient tracking error: $\mathbb{E}\|\mathbf{w}^k - B^\infty\mathbf{w}^k\|^2$;

where $\widehat{\mathbf{x}}^k := \boldsymbol{\pi}_r^\top\mathbf{x}^k$ and $\mathbf{t}^k := \sum_{i=1}^n(\frac{1}{m_i}\sum_{j=1}^{m_i}\|\mathbf{v}_{i,j}^k - \mathbf{x}^*\|_2^2)$. We establish the linear convergence of **AB-SAGA** by showing that all of these error terms linearly decay to zero, leading to $\mathbf{x}_i^k \to \mathbf{x}^*$, at each node i. We next establish an LTI system that governs the convergence rate of **AB-SAGA** in the following Lemma.

Lemma 1. *Consider* **AB-SAGA** *under Assumptions 1 and 2. If* $\alpha \leq \min\left\{\frac{1}{35\ell\sqrt{h_r h_c}}, \frac{\mu}{288n\ell^2(\boldsymbol{\pi}_r^\top\boldsymbol{\pi}_c)}\right\}$, $c \geq \frac{\log(4n)}{\log(1/\sigma_A)}$, *and* $d \geq \frac{\log(4n)}{\log(1/\sigma_B)}$; *then* $\forall k > 0, \mathbf{u}^{k+1} \leq G_\alpha\mathbf{u}^k$, *where* $\mathbf{u}^k \in \mathbb{R}^4$ *and* $G_\alpha \in \mathbb{R}^{4\times4}$ *are defined as:*

$$\mathbf{u}^k := \begin{bmatrix} \mathbb{E}[\|\mathbf{x}^k - A^\infty\mathbf{x}^k\|_{\boldsymbol{\pi}_r}^2] \\ \mathbb{E}[n\|\widehat{\mathbf{x}}^k - \mathbf{x}^*\|_2^2] \\ \mathbb{E}[\mathbf{t}^k] \\ \mathbb{E}[\ell^{-2}\|\mathbf{w}^k - B^\infty\mathbf{w}^k\|_{\boldsymbol{\pi}_c}^2] \end{bmatrix}, \tag{2}$$

$$G_\alpha := \begin{bmatrix} \frac{3}{4} & \alpha^2 g_1\sigma_A^{2c} & \alpha^2 g_2\sigma_A^{2c} & \alpha^2 g_3\sigma_A^{2c} \\ \alpha g_4 & 1-\alpha g_5 & \alpha^2 g_6 & \alpha g_7 \\ \frac{2}{m\underline{\pi}_r} & \frac{2}{m} & 1-\frac{1}{M} & 0 \\ \frac{146n\sigma_B^{2d}}{(1-\sigma_B^{2d})\underline{\pi}_r\underline{\pi}_c} & \frac{97n\sigma_B^{2d}}{(1-\sigma_B^{2d})\underline{\pi}_c} & \frac{26\sigma_B^{2d}}{(1-\sigma_B^{2d})\underline{\pi}_c} & \frac{3}{4} \end{bmatrix};$$

and the corresponding constants are given by

$$g_1 := \tfrac{40\ell^2 n\|\boldsymbol{\pi}_c\|_2^2\overline{\pi}_r}{1-\sigma_A^{2c}}, \quad g_2 := \tfrac{16\ell^2\|\boldsymbol{\pi}_c\|_2^2\overline{\pi}_r}{1-\sigma_A^{2c}}, \quad g_3 := \tfrac{8\ell^2\overline{\pi}_r\overline{\pi}_c}{1-\sigma_A^{2c}},$$
$$g_4 := \tfrac{8\ell^2 n\boldsymbol{\pi}_r^\top\boldsymbol{\pi}_c}{\mu\underline{\pi}_r}, \quad g_5 := \tfrac{\mu n\boldsymbol{\pi}_r^\top\boldsymbol{\pi}_c}{4},$$
$$g_6 := 3\ell^2 n(\boldsymbol{\pi}_r^\top\boldsymbol{\pi}_c)^2, \quad g_7 := \tfrac{5\ell^2\|\boldsymbol{\pi}_r\|_2^2\overline{\pi}_c}{\mu\boldsymbol{\pi}_r^\top\boldsymbol{\pi}_c}.$$

The proof of Lemma 1 is standard and follows similar procedures as in [14, 22]. With the help of this lemma, we next prove Theorem 1 based on the following key result.

Lemma 2. *[21] Let* $A \in \mathbb{R}^{n\times n}$ *be a non-negative matrix and* $\mathbf{x} \in \mathbb{R}^n$ *be a positive vector. If* $A\mathbf{x} \leq \beta\mathbf{x}$, *for some* $\beta > 0$, *then* $\rho(A) \leq \|\!|A|\!\|_\infty^{\mathbf{x}} \leq \beta$, *where* $\|\!|A|\!\|_\infty^{\mathbf{x}}$ *is the matrix norm induced by the weighted max-norm* $\|\cdot\|_\infty^{\mathbf{x}}$, *given that* $\mathbf{x} > \mathbf{0}_n$.

Proof of Theorem 1: First note that the system matrix G_α is non-negative. From Lemma 2, if there exists a positive vector $\boldsymbol{\delta}$ and a constant γ, such that $G_\alpha\boldsymbol{\delta} \leq \gamma\boldsymbol{\delta}$ element-wise, then $\rho(G_\alpha) \leq \|\!|G_\alpha|\!\|_\infty^{\boldsymbol{\delta}} \leq \gamma$. To this aim, let $\boldsymbol{\delta}$ have all positive elements $[\delta_1\ \delta_2\ \delta_3\ \delta_4]^\top$ and set $\gamma = \left(1 - \frac{1}{2}\alpha g_5\right)$, then the following set of inequalities must hold:

$$\frac{\alpha g_5}{2} + \frac{\sigma_A^{2c}}{\delta_1}\left(\alpha^2 g_1\delta_2 + \alpha^2 g_2\delta_3 + \alpha^2 g_3\delta_4\right) \leq \frac{1}{4}, \tag{3}$$

$$\alpha g_6 \leq \frac{g_5}{2}\frac{\delta_2}{\delta_3} - g_4\frac{\delta_1}{\delta_3} - g_7\frac{\delta_4}{\delta_3}, \tag{4}$$

$$\frac{\alpha g_5}{2} \leq \frac{1}{M} - \frac{2\underline{\pi}_r^{-1}}{m}\frac{\delta_1}{\delta_3} - \frac{2}{m}\frac{\delta_2}{\delta_3}, \tag{5}$$

$$\frac{\alpha g_5}{2} \leq \frac{1}{4} - \frac{\sigma_B^{2d}\left(146n\underline{\pi}_r^{-1}\delta_1 + 97n\delta_2 - 26\delta_3\right)}{\delta_4\underline{\pi}_c(1-\sigma_B^{2d})}. \tag{6}$$

We note that (4), (5) and (6) are valid for a range of stepsizes and communication rounds c and d, when their right hand sides are positive. To this aim, we first fix the elements of $\boldsymbol{\delta}$ independent of the stepsize and then find the bounds on α, i.e., $\delta_1 = 1$, $\delta_2 = \frac{64\tau_2\kappa^2}{\underline{\pi}_r}$, $\delta_3 = \frac{130\tau_2\kappa^2 M}{m\underline{\pi}_r}$ and $\delta_4 = \frac{40000n\kappa^2 M\underline{\pi}_r^{-1}\underline{\pi}_c^{-1}}{m\tau_1(1-\sigma_B^{2d})}$, when $\sigma_B^d < \frac{\boldsymbol{\pi}_r^\top\boldsymbol{\pi}_c}{201\kappa}\sqrt{\frac{m}{nMh_c}}$ and $\tau_1 := 1 - \frac{40000n\sigma_B^{2d}\kappa^2 Mh_c}{m(1-\sigma_B^{2d})(\boldsymbol{\pi}_r^\top\boldsymbol{\pi}_c)^2}, \tau_2 := 1 + \frac{40000n\kappa^2 Mh_c}{(\boldsymbol{\pi}_r^\top\boldsymbol{\pi}_c)^2\tau_1 m(1-\sigma_B^{2d})}$.

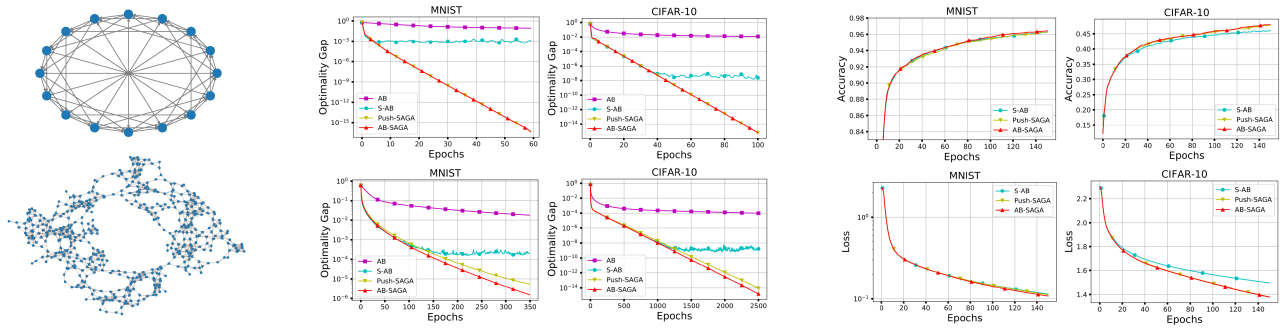

Fig. 1. (Left) Directed (doubly stochastic) exponential graph (top) with $n = 16$ nodes and directed geometric graph (bottom) with $n = 500$ nodes. (Center) Optimality gap for logistic regression classifier trained over directed exponential graph (top) and directed geometric graph (bottom); (right) Test accuracy and training loss for neural networks trained over a geometric graph.

It can be verified that for these values of $\boldsymbol{\delta}$, the right hand sides of (4), (5) and (6) are positive. We next solve for finding a range for the stepsize α. From (4), we have

$$\alpha < \frac{g_5}{2g_6}\frac{\delta_2}{\delta_3} - \frac{g_4}{g_6}\frac{\delta_1}{\delta_3} - \frac{g_7}{g_6}\frac{\delta_4}{\delta_3}, \iff \alpha \leq \frac{m}{135M\kappa\ell(\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c)}.$$

Similarly, plugging these values of $\boldsymbol{\delta}$, in (5) yields

$$\alpha \leq \frac{2}{Mg_5} - \frac{4\underline{\pi}_r^{-1}}{mg_5}\frac{\delta_1}{\delta_3} - \frac{4}{mg_5}\frac{\delta_2}{\delta_3}, \iff \alpha \leq \frac{1}{9M\mu(\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c)}$$

To find a bound on α from (6), we need to ensure that $\sigma_B^d \leq \frac{\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c}{1265\kappa}\sqrt{\frac{m}{nMh_c}}$ and therefore, we have

$$\alpha \leq \frac{1}{2g_5} - \frac{\sigma_B^{2d}\left(292n\underline{\pi}_r^{-1}\delta_1 + 194n\delta_2 - 52\delta_3\right)}{g_5\delta_4\underline{\pi}_c(1-\sigma_B^{2d})}$$

$$\iff \alpha \leq \frac{1}{225\mu}\left(\frac{1}{n(\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c)}\right).$$

We note that (3) has solution if we bound $\alpha \leq \frac{1}{\mu}\cdot\frac{2}{5n(\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c)}$ for the first term, $\alpha \leq \frac{1}{\kappa\ell}$ for the rest of the terms and ensure

$$\sigma_A^{2c} < \min\left\{\frac{m(1-\sigma_A^{2c})}{51200nM\tau_2 h_r}, \frac{m\tau_1(1-\sigma_A^{2c})(1-\sigma_B^{2d})}{640000nMh_r h_c}\right\}.$$

To simplify the bounds on σ_A^c and σ_B^d, it can be verified that $\sigma_A^c < \frac{m(1-\sigma_B^{2d})}{90512nM\kappa}\sqrt{\frac{\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c}{h_r h_c}}$ and $\sigma_B^d < \frac{\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c}{1265\kappa}\sqrt{\frac{m}{nMh_c}}$ satisfy all of the above bounds. We next define the smallest upper bound $\overline{\alpha}$ on the stepsize as

$$\overline{\alpha} := \min\left\{\frac{1}{35\ell\sqrt{h_r h_c}}, \frac{m}{288Mn\kappa\ell(\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c)}, \frac{1}{9\mu M(\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c)}\right\}.$$

If $\alpha \in (0, \overline{\alpha}]$, and the communication rounds,

$$c > \frac{\log\left(\frac{m(1-\sigma_B^{2d})}{90512nM\kappa}\sqrt{\frac{\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c}{h_r h_c}}\right)}{\log \sigma_A}, \quad d > \frac{\log\left(\frac{\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c}{1265\kappa}\sqrt{\frac{m}{nMh_c}}\right)}{\log \sigma_B};$$

from Lemma 2, the spectral radius $\rho(G_\alpha) \leq \gamma = 1 - \frac{\alpha\mu n\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c}{8}$. Furthermore, if $\alpha = \overline{\alpha}$ and $\psi := \frac{\sqrt{h_r h_c}}{n(\boldsymbol{\pi}_r^\top \boldsymbol{\pi}_c)}$,

$$\rho(G_\alpha) \leq 1 - \min\left\{\frac{1}{35\kappa\psi}, \frac{m}{288\kappa^2 M}, \frac{1}{9M}\right\}.$$

and the theorem follows. □

5. NUMERICAL EXPERIMENTS

In this section, we illustrate the performance of **AB-SAGA** and compare it with the existing methods for directed graphs, i.e., **AB**, **S-AB**, and **Push-SAGA** [4, 14, 22].

Logistic Regression: We consider binary classification, using logistic regression with a strongly convex regularizer, for $N = 12{,}000$ labeled images, taken from the MNIST and CIFAR-10 datasets, and distributed over strongly connected directed exponential and geometric graphs. Fig. 1 plots the optimality gaps $F(\overline{\mathbf{x}}^k) - F(\mathbf{x}^*)$ versus the number of epochs, where $\overline{\mathbf{x}}^k := \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i^k$. We note that one epoch is m_i updates for stochastic methods and a single update for **AB**. It can be seen, in Fig. 1 (center), that **AB-SAGA** converges linearly to the optimal solution outperforming all other methods. We note that **Push-SAGA** is potentially slower because it requires additional iterations for eigenvector estimation.

Neural Networks: We next consider multi-class classification over $N = 60{,}000$ (MNIST and CIFAR-10). Each node trains a local neural network with a hidden layer of 64 neurons and a fully connected output layer of 10 neurons. We plot the training loss $F(\overline{\mathbf{x}}^k)$ and test accuracy for the stochastic methods: **S-AB**, **Push-SAGA** and **AB-SAGA** in Fig. 1 (right), trained over a directed geometric graph ($n = 500$). It can be observed that **AB-SAGA** achieves a lower loss and an improved test accuracy over the other methods.

6. CONCLUSIONS

This paper describes a first-order stochastic method to minimize a distributed optimization problem over strongly connected directed graphs. We show linear convergence of the proposed method **AB-SAGA** to the optimal solution under weaker assumptions compared to the earlier work. Our results provide a key insight by describing a directivity constant that quantifies the directed nature of the communication network and we further show linear speed-up of **AB-SAGA** as compared to its centralized counterpart. Numerical experiments illustrate the convergence guarantees for strongly convex regression problems and non-convex neural networks.

7. REFERENCES

[1] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM Review*, vol. 60, no. 2, pp. 223–311, 2018.

[2] D. Driggs, J. Tang, J. Liang, M. Davies, and C.-B. Schönlieb, "A stochastic proximal alternating minimization for nonsmooth and nonconvex optimization," *SIAM Journal of Imaging Sciences*, vol. 14, no. 4, pp. 1932–1970, 2021.

[3] R. Xin, S. Kar, and U. A. Khan, "Decentralized stochastic optimization and machine learning," *IEEE Signal Processing Magazine*, vol. 3, pp. 102–113, May 2020.

[4] R. Xin, S. Pu, A. Nedić, and U. A. Khan, "A general framework for decentralized optimization with first-order methods," *Proceedings of the IEEE*, vol. 108, no. 11, pp. 1869–1889, 2020.

[5] Q. Song, D. Meng, and F. Liu, "Consensus-based iterative learning of heterogeneous agents with application to distributed optimization," *Automatica*, vol. 137, pp. 110096, 2022.

[6] 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.

[7] J. Chen and A. H. Sayed, "Diffusion adaptation strategies for distributed optimization and learning over networks," *IEEE Transactions on Signal Processing*, vol. 60, no. 8, pp. 4289–4305, Aug. 2012.

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

[9] C. N. Hadjicostis and T. Charalambous, "Average consensus in the presence of delays in directed graph topologies," *IEEE Transactions on Automatic Control*, vol. 59, no. 3, pp. 763–768, 2014.

[10] M. I. Qureshi, R. Xin, S. Kar, and U. A. Khan, "S-ADDOPT: Decentralized stochastic first-order optimization over directed graphs," *IEEE Control Systems Letters*, vol. 5, no. 3, pp. 953–958, 2021.

[11] P. Di Lorenzo and G. Scutari, "NEXT: In-network nonconvex optimization," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 2, no. 2, pp. 120–136, 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] C. Xi, R. Xin, and U. A. Khan, "ADD-OPT: Accelerated distributed directed optimization," *IEEE Transactions on Automatic Control*, vol. 63, no. 5, pp. 1329–1339, 2017.

[14] M. I. Qureshi, R. Xin, S. Kar, and U. A. Khan, "Push-SAGA: A decentralized stochastic algorithm with variance reduction over directed graphs," *IEEE Control Systems Letters*, 2022.

[15] A. Defazio, F. Bach, and S. Lacoste-Julien, "SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives," in *Advances in Neural Information Processing Systems*, 2014, pp. 1646–1654.

[16] A. Olshevsky, "Linear time average consensus and distributed optimization on fixed graphs," *SIAM Journal on Control and Optimization*, vol. 55, no. 6, pp. 3990–4014, 2017.

[17] S. Pu and A. Nedić, "Distributed stochastic gradient tracking methods," *Mathematical Programming*, 2020.

[18] R. Xin, U. A. Khan, and S. Kar, "Variance-reduced decentralized stochastic optimization with accelerated convergence," *IEEE Transactions on Signal Processing*, vol. 68, pp. 6255–6271, 2020.

[19] R. Xin, U. A. Khan, and S. Kar, "A fast randomized incremental gradient method for decentralized nonconvex optimization," *IEEE Transactions on Automatic Control*, vol. 67, no. 10, pp. 5150–5165, 2022.

[20] R. Xin, U. A. Khan, and S. Kar, "Fast decentralized nonconvex finite-sum optimization with recursive variane reduction," *SIAM Journal on Optimization*, vol. 32, no. 1, pp. 1–28, 2022.

[21] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, 2nd edition, 2012.

[22] R. Xin, A. K. Sahu, U. A. Khan, , and S. Kar, "Distributed stochastic optimization with gradient tracking over strongly-connected networks," in *58th IEEE Conference of Decision and Control*, Nice, France, 2019.

[23] R. Xin, U. A. Khan, and S. Kar, "An improved convergence analysis for decentralized online stochastic nonconvex optimization," *IEEE Transactions on Signal Processing*, 2021.

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