S-ADDOPT: Decentralized stochastic first-order optimization over directed graphs

Muhammad I. Qureshi†, Ran Xin‡, Soummya Kar‡, and Usman A. Khan†
†Tufts University, Medford, MA, USA, ‡Carnegie Mellon University, Pittsburgh, PA, USA

Abstract

In this report, we study decentralized stochastic optimization to minimize a sum of smooth and strongly convex cost functions when the functions are distributed over a directed network of nodes. In contrast to the existing work, we use gradient tracking to improve certain aspects of the resulting algorithm. In particular, we propose the **S-ADDOPT** algorithm that assumes a stochastic first-order oracle at each node and show that for a constant step-size α, each node converges linearly inside an error ball around the optimal solution, the size of which is controlled by α. For decaying step-sizes $\mathcal{O}(1/k)$, we show that **S-ADDOPT** reaches the exact solution sublinearly at $\mathcal{O}(1/k)$ and its convergence is asymptotically network-independent. Thus the asymptotic behavior of **S-ADDOPT** is comparable to the centralized stochastic gradient descent. Numerical experiments over both strongly convex and non-convex problems illustrate the convergence behavior and the performance comparison of the proposed algorithm.

I. Introduction

This report considers minimizing a sum of smooth and strongly convex functions $F(\mathbf{z}) = \sum_{i=1}^{n} f_i(\mathbf{z})$ over a network of n nodes. We assume that each f_i is private to only on node i and that the nodes communicate over a directed graph (digraph) to solve the underlying problem. Such problems have found significant applications traditionally in the areas of signal processing and control [1], [2] and more recently in machine learning problems [3]–[6]. Gradient descent (**GD**) is one of the simplest algorithms for function minimization and requires the true gradient ∇F. When this information is not available, **GD** is implemented with stochastic gradients and the resulting method is called stochastic gradient descent (**SGD**). As the data becomes large-scale and geographically diverse, **GD** and **SGD** present storage and communication challenges. In such cases, decentralized methods are attractive as they are locally implemented and rely on communication among nearby nodes.

Related work on decentralized first-order methods can be found in [7]–[12]. Of relevance is Distributed Gradient Descent (**DGD**) that converges sublinearly to the optimal solution with decaying step-sizes [7] and linearly to an inexact solution with a constant step-size [8]. Its stochastic variant **DSGD** can be found in [9], [10], which is further extended with the help of gradient tracking [13]–[15] in [12] where inexact linear convergence in addition

The authors acknowledge the support of NSF under awards CCF-1513936, CMMI-1903972, and CBET-1935555.

to asymptotic network independence are shown; see also [16]–[18] and references therein. More recently, variance reduction has been used to show linear convergence for smooth and strongly convex finite-sum problems [11]. However, all of these decentralized stochastic algorithms are built on undirected graphs, see [19] for a friendly tutorial. Related work on directed graphs includes [14], [15], [20]–[24] where true gradients are used, and [16], [25]–[27] on stochastic methods, all of which use the push-sum algorithm [28] to achieve agreement with an exception of [15], [27], [29], [30] that employ updates with both row and column stochastic weights to avoid the eigenvector estimation in push-sum.

In this report, we present **S-ADDOPT** for decentralized stochastic optimization over directed graphs. In particular, **S-ADDOPT** adds gradient tracking to **SGP** (stochastic gradient push) [16], [25], [26] and can be viewed as a stochastic extension of **ADDOPT** [14], [31] that uses true gradients. Of significant relevance is [12] that is applicable to undirected graphs and is based on doubly stochastic weights. Since **S-ADDOPT** is based on directed graphs, it essentially extends the algorithm in [12] with the help of push-sum when the network weights are restricted to be column stochastic. A similar algorithm based on row-stochastic weights is also immediate by apply the extension and analysis in this report to FROST [23], [24].

The main contributions of this report are as follows: (i) We develop a stochastic algorithm over directed graphs by combining push-sum with gradient tacking; (ii) For a constant step-size α, we show that each node converges linearly inside an error ball around the optimal solution, and further show that the size of the error ball is controlled by α. (iii) For decaying step-sizes $\mathcal{O}(1/k)$, we show that **S-ADDOPT** is asymptotically network-independent and reaches the exact solution sublinearly at $\mathcal{O}(1/k)$, while the network agreement error decays at a faster rate of $\mathcal{O}(1/k^2)$. (iv) We explicitly quantify the directed nature of graphs using a directivity constant τ, which makes this work a generalization of **DSGD**, **SGP**, and the method proposed in [12]. The directivity constant τ is 1 for undirected graphs and thus the results apply to undirected graphs as a special case. The rest of this report is organized as follows. We formalize the optimization problem, list the underlying assumptions, and describe **S-ADDOPT** in Section II. We then present the main results in Section III and the convergence analysis in Section IV. Finally, we provide numerical experiments in Section V and conclude the report in Section VI.

Basic Notation: We use uppercase italic letters for matrices and lowercase bold letters for vectors. We use I_n for the $n \times n$ identity matrix and $\mathbf{1}_n$ denotes the column vector of n ones. A column stochastic matrix is such that it is non-negative and all of its columns sum to 1. For a primitive column stochastic matrix $\underline{B} \in \mathbb{R}^{n\times n}$, we have $\underline{B}^\infty = \boldsymbol{\pi}\mathbf{1}_n^\top$, from the Perron-Frobenius theorem [32], where $\boldsymbol{\pi}$ and $\mathbf{1}_n^\top$ are its right and left Perron eigenvectors. For a matrix G, $\rho(G)$ is its spectral radius. We denote the Euclidean (vector) norm by $\|\cdot\|_2$ and define a weighted inner product as $\langle \mathbf{x}, \mathbf{y} \rangle_{\boldsymbol{\pi}} := \mathbf{x}^\top \text{diag}(\boldsymbol{\pi})^{-1}\mathbf{y}$, for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$, which leads to a weighted Euclidean norm: $\|\mathbf{x}\|_{\boldsymbol{\pi}} := \|\text{diag}(\sqrt{\boldsymbol{\pi}})^{-1}\mathbf{x}\|_2$. We denote $|||\cdot|||_{\boldsymbol{\pi}}$ as the matrix norm induced by $\|\cdot\|_{\boldsymbol{\pi}}$ such that $\forall X \in \mathbb{R}^{n\times n}$, $|||\,X\,||| := \left|\left|\left|\,\text{diag}(\sqrt{\boldsymbol{\pi}})^{-1}\ X\ \text{diag}(\sqrt{\boldsymbol{\pi}})\,\right|\right|\right|_2$. Note that these norms are related as $\|\cdot\|_{\boldsymbol{\pi}} \leq \underline{\pi}^{-0.5}\|\cdot\|_2$ and $\|\cdot\|_2 \leq \overline{\pi}^{0.5}\|\cdot\|_{\boldsymbol{\pi}}$, where $\overline{\pi}$ and $\underline{\pi}$ are the maximum and minimum elements in $\boldsymbol{\pi}$, while $|||\,\underline{B}\,|||_{\boldsymbol{\pi}} = |||\,\underline{B}^\infty\,|||_{\boldsymbol{\pi}} = |||\,I_n - \underline{B}^\infty\,|||_{\boldsymbol{\pi}} = 1$. Finally, it is shown in [27] that $\sigma_B := |||\,\underline{B} - \underline{B}^\infty\,|||_{\boldsymbol{\pi}} < 1$.

II. Problem Formulation

Consider n nodes communicating over a strongly-connected directed graph (digraph), $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, 2, 3, \ldots, n\}$ is the set of agents and $\mathcal{E}$ is the collection of ordered pairs, $(i, j), i, j \in \mathcal{V}$, such that node i receives information from node j. We let $\mathcal{N}_i^{\text{out}}$ (resp. $\mathcal{N}_i^{\text{in}}$) to denote the set of out-neighbors (resp. in-neighbors) of node i, i.e., nodes that can receive information from i, and $|\mathcal{N}_i^{\text{out}}|$ is the out-degree of node i. Note that both $\mathcal{N}_i^{\text{out}}$ and $\mathcal{N}_i^{\text{in}}$ include node i. The nodes collaborate to solve the following optimization problem:

$$\mathbf{P}: \qquad \min_{\mathbf{z} \in \mathbb{R}^p} F(\mathbf{z}) := \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{z}),$$

where each node i possesses a private cost function $f_i : \mathbb{R}^p \to \mathbb{R}$. We make the following assumptions.

Assumption 1. *The communication graph $\mathcal{G}$ is a strongly-connected directed graph and each node has the knowledge of its out-degree $|\mathcal{N}_i^{out}|$.*

Assumption 2. *Each local cost function f_i (and thus F) is μ-strongly convex and ℓ-smooth, i.e., $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^p$ and $\forall i \in \mathcal{V}$, there exist positive constants μ and ℓ such that*

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

Note that the ratio $\kappa := \frac{\ell}{\mu}$ is called the condition number of the function f_i. We have that $\ell \geq \mu$ and thus $\kappa \geq 1$.

Assumption 3. *Each node has access to a stochastic first-order oracle $\texttt{SFO}$ that returns a stochastic gradient $\nabla \widehat{f}_i(\mathbf{z}_k^i)$ for any $\mathbf{z}_k^i \in \mathbb{R}^p$ such that*

$$\mathbb{E}\left[\nabla \widehat{f}_i(\mathbf{z}_k^i) | \mathbf{z}_k^i\right] = \nabla f_i(\mathbf{z}_k^i),$$
$$\mathbb{E}\left[\|\nabla \widehat{f}_i(\mathbf{z}_k^i) - \nabla f_i(\mathbf{z}_k^i)\|_2^2 | \mathbf{z}_k^i\right] \leq \sigma^2.$$

These assumptions are standard in the related literature. The bounded variance assumption however can be relaxed, see [6], for example. Due to Assumption 2, we note that F has a unique minimizer that is denoted by $\mathbf{z}^*$. The proposed algorithm to solve Problem $\mathbf{P}$ is described next.

A. **S-ADDOPT**: *Algorithm*

The **S-ADDOPT** algorithm to solve Problem $\mathbf{P}$ is formally described in Algorithm 1. We note that the set of weights $\underline{B} = \{b_{ij}\}$ is such that $\underline{B}$ is column stochastic. A valid choice is $b_{ji} = |\mathcal{N}_i^{\text{out}}|^{-1}$, for each $j \in \mathcal{N}_i^{\text{out}}$ and zero otherwise, recall Assumption 1. Each agent i maintains three state vectors, i.e., $\mathbf{x}_k^i, \mathbf{w}_k^i, \mathbf{z}_k^i \in \mathbb{R}^p$ and a (positive) scalar y_k^i at each iteration k. The first update $\mathbf{x}_{k+1}^i$ is similar to **DSGD**, where the stochastic gradient $\nabla \widehat{f}_i(\mathbf{x}_k^i)$ is replaced with $\mathbf{w}_k^i$. This auxiliary variable $\mathbf{w}_k^i$ is based on dynamic average-consensus [33] and in fact tracks the global gradient ∇F when viewed as a non-stochastic update (see [13]–[15], [34] for details). However, since the weight matrix $\underline{B}$ is not row-stochastic, the variables $\mathbf{x}_k^i$'s do not agree on a solution and converge with a certain imbalance that is due to the fact that $\mathbf{1}_n$ is not the right Perron eigenvector of $\underline{B}$. This imbalance is canceled in the $\mathbf{z}_k^i$-update with the help of a scaling by y_k^i, since y_k^i estimates the i-th component of $\boldsymbol{\pi}$ (recall that $\underline{B}\boldsymbol{\pi} = \boldsymbol{\pi}$). We note that **`S-ADDOPT`** is in fact a stochastic extension of **`ADDOPT`**, where true local gradients ∇f_i's are used at each node.

Algorithm 1 **`S-ADDOPT`**: At each node i

Require: $\mathbf{x}_0^i \in \mathbb{R}^p, \mathbf{z}_0^i = \mathbf{x}_0^i, y_0^i = 1, \mathbf{w}_0^i = \nabla \widehat{f}_i(\mathbf{z}_0^i), \alpha > 0$

for $k = 0, 1, 2, \cdots$ **do**

  State update: $\mathbf{x}_{k+1}^i = \sum_{j=1}^n b_{ij}\mathbf{x}_k^j - \alpha \mathbf{w}_k^i$

  Eigenvector est.: $y_{k+1}^i = \sum_{j=1}^n b_{ij} y_k^j$

  Push-sum update: $\mathbf{z}_{k+1}^i = \mathbf{x}_{k+1}^i / y_{k+1}^i$

  Gradient tracking update: $\mathbf{w}_{k+1}^i = \sum_{j=1}^n b_{ij}\mathbf{w}_k^j + \nabla \widehat{f}_i(\mathbf{z}_{k+1}^i) - \nabla \widehat{f}_i(\mathbf{z}_k^i)$

end for

`S-ADDOPT` can be compactly written in a vector form with the help of the following notation. Let $\mathbf{x}_k, \mathbf{z}_k, \mathbf{w}_k$, all in $\mathbb{R}^{np}$ concatenate the local states $\mathbf{x}_k^i, \mathbf{z}_k^i, \mathbf{w}_k^i$ (all in $\mathbb{R}^p$) at the nodes and $\mathbf{y}_k \in \mathbb{R}^n$ stacks the y_k^i's. Let $\otimes$ denote the Kronecker product and define $B := \underline{B} \otimes I_p$, and let $Y_k := \text{diag}(\mathbf{y}_k) \otimes I_p$. Then **`S-ADDOPT`** described in Algorithm 1 can be written in a vector form as

$$\mathbf{x}_{k+1} = B\mathbf{x}_k - \alpha \mathbf{w}_k, \tag{1a}$$

$$\mathbf{y}_{k+1} = \underline{B}\mathbf{y}_k, \tag{1b}$$

$$\mathbf{z}_{k+1} = Y_{k+1}^{-1}\mathbf{x}_{k+1}, \tag{1c}$$

$$\mathbf{w}_{k+1} = B\mathbf{w}_k + \nabla \widehat{f}(\mathbf{z}_{k+1}) - \nabla \widehat{f}(\mathbf{z}_k). \tag{1d}$$

In the following sections, we summarize the main results (Section III) and provide the convergence analysis (Section IV) of **`S-ADDOPT`**. Subsequently, we compare its performance with related algorithms on digraphs in Section V.

III. Main Results

We use $p = 1$ for simplicity and thus $B = \underline{B}$. Before we proceed, we define $\overline{\mathbf{x}}_k := \frac{1}{n}\mathbf{1}_n^\top \mathbf{x}_k$, and $\widehat{\mathbf{x}}_k := B^\infty \mathbf{x}_k$, which are the mean and weighted averages of $\mathbf{x}_k^i$'s, respectively, and $y := \sup_k \left\|\left| Y_k \right|\right\|_2$, $y_- := \sup_k \left\|\left| Y_k^{-1} \right|\right\|_2$. We next provide two useful lemmas.

Lemma 1. *[14], [27] Consider Assumption* 1 *and define* $Y^\infty := \lim_{k\to\infty} Y_k$, $h := \overline{\pi}/\underline{\pi}$, *and* $\beta := \sqrt{h}\|\mathbf{1}_n - n\boldsymbol{\pi}\|_2$. *Then* $\left\|\left| Y_k - Y^\infty \right|\right\|_2 \leq \beta\sigma_B^k$, $\forall k \geq 0$.

Proof. Note that $\forall k \geq 0$, $\mathbf{y}_\infty = B^\infty \mathbf{y}_k$. Thus we have

$$\left\|\left| Y_k - Y^\infty \right|\right\|_2 \leq \|\mathbf{y}_k - \mathbf{y}_\infty\|_2 \leq \sqrt{\overline{\pi}} \left\|\left| B - B^\infty \right|\right\|_{\boldsymbol{\pi}} \|\mathbf{y}_{k-1} - \mathbf{y}_\infty\|_{\boldsymbol{\pi}} \leq \sigma_B^k \sqrt{h}\|\mathbf{y}_0 - \mathbf{y}_\infty\|_2.$$

and the proof follows. □

Lemma 2. *Define* $\mathbf{e}_k := \frac{1}{n}\mathbb{E}[\|\mathbf{z}_k - \mathbf{1}_n\mathbf{z}^*\|_2^2]$ *as the mean error in the network. We have*

$$\mathbf{e}_k \leq \frac{\omega}{n}\mathbb{E}\|\mathbf{x}_k - \widehat{\mathbf{x}}_k\|_{\boldsymbol{\pi}}^2 + \omega\beta^2\sigma_B^k\|\mathbf{z}^*\|_2^2 + \omega y^2\mathbb{E}\|\overline{\mathbf{x}}_k - \mathbf{z}^*\|_{\boldsymbol{\pi}}^2, \tag{2}$$

$$\mathbf{e}_k \leq \psi\mathbb{E}\|\mathbf{x}_k - \widehat{\mathbf{x}}_k\|_{\boldsymbol{\pi}}^2 + \psi\beta\sigma_B^k\mathbb{E}\|\mathbf{x}_k\|_{\boldsymbol{\pi}}^2 + 2\mathbb{E}\|\overline{\mathbf{x}}_k - \mathbf{z}^*\|_2^2, \tag{3}$$

where $\omega := 3y_-^2\overline{\pi}$ *and* $\psi := 2y_-^2\overline{\pi}(1+\beta)/n$*.*

We now provide the main results on **S-ADDOPT**.

Theorem 1. *Let Assumptions 1, 2, and 3 hold and let the step-size* α *be a constant such that,*

$$\alpha \leq \frac{1}{\ell\sqrt{\kappa}} \cdot \frac{(1-\sigma_B^2)^2}{51\sqrt{\tau}}, \tag{4}$$

where $\tau := y_-^6 y^2 h(1+\beta)$ *is the directivity constant. Then* $\mathbf{e}_k$ *converges linearly, at a rate* $\gamma, \gamma \in [0,1)$*, to a ball around* $\mathbf{z}^*$*, i.e.,*

$$\limsup_{k\to\infty} \mathbf{e}_k = \alpha\,\mathcal{O}\left(\frac{\sigma^2}{n\mu}\right) + \alpha^2\,\mathcal{O}\left(\frac{\ell^2\sigma^2}{\mu^2(1-\sigma_B^2)^4}\right). \tag{5}$$

The proof of Theorem 1 is provided in the next Section. It essentially shows that **S-ADDOPT** converges linearly with a constant step-size to an error ball around $\mathbf{z}^*$, the size of which however is controlled by α. We note that $\tau \geq 1$ can be considered as a directivity constant and is large when the graph is more directed as quantified by e.g., the constant h (in addition to the other constants in τ); clearly, for undirected graphs $\tau = 1$ and thus Theorem 1 is applicable to undirected graphs as a special case. We further note that the first term in (5) is due to the variance σ^2 of the stochastic gradients and does not have a network dependence, i.e., a scaling with $(1-\sigma_B^2)^{-1}$. The rate of convergence of **S-ADDOPT** thus is comparable to the **SGD** (up to some constant factors) when the step-size α is sufficiently small since the second term has a higher order of α. The result in Theorem 1 is similar to what was obtained for undirected graphs in [12], where the network dependence is $\mathcal{O}((1-\sigma_B^2)^{-3})$. We next provide an upper bound on the linear rate γ.

Corollary 1. *Let Assumptions 1, 2 and 3 hold. If the step-size follows* $\alpha \leq \frac{3}{40}\left(\frac{1-\sigma_B^2}{\mu}\right)$*, then the linear rate parameter* γ *in Theorem 1 is such that*

$$\gamma \leq 1 - \frac{\alpha\mu}{3}.$$

The proof of Corollary 1 is available in Appendix B and follows the same arguments as in [12]. Going back to Theorem 1, note that the exact expression of (5) is provided later in the convergence analysis, see (15), where we dropped the higher powers of α when writing (5). We note from (15) that all terms in the residual are a function of σ^2 and thus **S-ADDOPT** recovers the exact linear convergence as σ^2 vanishes. When σ^2 is not zero, exact convergence is achievable albeit at a sublinear rate with decaying step-sizes. We provide this result below.

Theorem 2. *Let Assumptions 1, 2, and 3 hold. Consider* **S-ADDOPT** *with decaying step-sizes* $\alpha_k := \frac{\theta}{m+k}, \theta > \frac{1}{\mu}$ *and* m *such that*

$$\begin{cases} m > \max\left\{\frac{\theta(\ell+\mu)}{2}, \frac{6\ell\theta y_- \sqrt{(1+\sigma_B^2)h}}{1-\sigma_B^2}\right\}, \\ \frac{(1-\sigma_B^2)^2}{6\theta^2(1+\sigma_B^2)}\left(\frac{1-\sigma_B^2}{2} - \frac{2m+1}{(m+1)^2}\right) > \frac{E_2}{m^2} + \left(\frac{\theta^3\ell^6 E_1 E_3}{m^4 n(\theta\mu-1)}\right)\left(\frac{\theta\mu+m}{m\mu}\right), \end{cases}$$

for some constants E_1, E_2, E_3. *Select* $\widetilde{S}$ *large enough such that* $\forall k \geq \widetilde{S}, \sigma_B^k \leq \frac{1}{n(m+k)^2}$, *then we have*

$$\mathbb{E}\|\mathbf{x}_k - \widehat{\mathbf{x}}_k\|_{\boldsymbol{\pi}}^2 \leq \frac{\mathcal{O}(1)}{(m+k)^2},$$

$$\mathbb{E}\|\overline{\mathbf{x}}_k - \mathbf{z}^*\|_2^2 \leq \frac{2\theta^2\sigma^2}{n(\theta\mu-1)(m+k)} + \frac{\mathcal{O}(1)}{(m+k)^{\theta\mu}} + \frac{\mathcal{O}(1)}{(m+k)^2},$$

which leads to $\mathbf{e}_k \to 0$ *at a network-independent convergence rate of* $\mathcal{O}(\frac{1}{k})$.

Theorem 2, formally analyzed in the next section, shows that the error $\mathbf{e}_k$ in **S-ADDOPT** asymptotically converges to the exact solution at a rate dominant by $\frac{4\theta^2\sigma^2}{n(\theta\mu-1)k}$, which is network-independent since all other terms decay faster, and thus **S-ADDOPT** matches the rate of **SGD** (up to some constant factors); see also [12], [16]–[18]. It can also be verified that the network reaches an agreement at $\mathcal{O}(1/k^2)$.

IV. Convergence Analysis

To aid the analysis of Theorems 1 and 2, we first develop a dynamical system that characterizes **S-ADDOPT** for both constant and decaying step-sizes. We find inter-relationships between the following three terms:

(i) Network agreement error, $\mathbb{E}\|\mathbf{x}_k - B^\infty\mathbf{x}_k\|_{\boldsymbol{\pi}}^2$,

(ii) Optimality gap, $\mathbb{E}\|\overline{\mathbf{x}}_k - \mathbf{z}^*\|_2^2$,

(iii) Gradient tracking error, $\mathbb{E}\|\mathbf{w}_k - B^\infty\mathbf{w}_k\|_{\boldsymbol{\pi}}^2$,

to write an LTI system of equations governing **S-ADDOPT**. For simplicity, we assume $p=1$. Denote $\mathbf{t}_k, \mathbf{s}_k, \mathbf{c} \in \mathbb{R}^3$, and $A_\alpha, H_k \in \mathbb{R}^{3\text{x}3}$ for all k as

$$\mathbf{t}_k := \begin{bmatrix} \mathbb{E}[\|\mathbf{x}_k - B^\infty\mathbf{x}_k\|_{\boldsymbol{\pi}}^2] \\ \mathbb{E}[\|\overline{\mathbf{x}}_k - \mathbf{z}^*\|_2^2] \\ \mathbb{E}[\|\mathbf{w}_k - B^\infty\mathbf{w}_k\|_{\boldsymbol{\pi}}^2] \end{bmatrix}, \quad \mathbf{s}_k := \begin{bmatrix} \mathbb{E}[\|\mathbf{x}_k\|_2^2] \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{c} := \begin{bmatrix} 0 \\ \alpha^2\frac{\sigma^2}{n} \\ C_\sigma \end{bmatrix},$$

$$H_k := \begin{bmatrix} 0 & 0 & 0 \\ h_1\sigma_B^k & 0 & 0 \\ (h_2+\alpha^2 h_3)\sigma_B^k & 0 & 0 \end{bmatrix}, \quad A_\alpha := \begin{bmatrix} \frac{1+\sigma_B^2}{2} & 0 & \alpha^2\frac{1+\sigma_B^2}{1-\sigma_B^2} \\ \alpha^2 g_1 + \alpha g_2 & 1-\alpha\mu & 0 \\ g_3 + \alpha^2 g_4 & \alpha^2 g_5 & \frac{5+\sigma_B^2}{6} \end{bmatrix}, \tag{6}$$

where the constants are defined as:

$$\begin{aligned}
&g_1 := \left(\frac{\ell^2 y_-^2}{n}\right)(1+\beta\sigma_B)\overline{\pi}, && g_2 := \left(\frac{\ell^2 y_-^2}{n\mu}\right)(1+\beta\sigma_B)\overline{\pi}, && g_3 := 4k_2, \\
&g_4 := 2\ell^2 y^2 k_2 k_3(1+\beta\sigma_B), && g_5 := 18\ell^4 q y_-^4 y^2 \underline{\pi}^{-1}, && k_1 := \tfrac{1-\sigma_B^2}{3}, \\
&C_\sigma := \sigma^2\left(c_1 + \alpha^2 c_2\right), && c_1 := 4qn\underline{\pi}^{-1}, && k_2 := 6\ell^2 q y_-^2 h \\
&c_2 := 12\ell^2 q y_-^4 y^2 k_3 \underline{\pi}^{-1}, && h_1 := y_-^2\beta\left(\frac{\alpha\ell^2}{\mu} + \alpha^2\ell^2\right)(\beta+1), && k_3 := \tfrac{2k_1 - 3k_2\alpha^2}{k_1 - 2k_2\alpha^2}, \\
&h_2 := 24\ell^2 q y_-^4 \beta^2 \underline{\pi}^{-1}, && h_3 := 12\ell^4 q y_-^6 y^2 k_3 \beta \underline{\pi}^{-1}(\beta+1), && q := \tfrac{1+\sigma_B^2}{1-\sigma_B^2}.
\end{aligned}$$

With $\alpha \leq \left(\frac{1-\sigma_B^2}{9\ell}\right)\frac{1}{y_-\sqrt{h}}$, we have that

$$\mathbf{t}_{k+1} \leq A_\alpha \mathbf{t}_k + H_k \mathbf{s}_k + \mathbf{c}. \tag{7}$$

The derivation of the above inequality is available in Appendix A. We now provide the proofs of Theorems 1 and 2.

A. *Proof of Theorem 1*

From [12] Lemma 5, for a 3×3 non-negative, irreducible matrix $A_\alpha = \{a_{ij}\}$ with $\{a_{ii}\} < \lambda^*$, we have $\rho(A_\alpha) < \lambda^*$ if and only if $\det(\lambda^* I_3 - A_\alpha) > 0$. For A_α in (6), $a_{11}, a_{33} < 1$ since $\sigma_B \in [0,1)$, and $a_{22} < 1$ since $\alpha < \frac{1}{\ell}$ and $\ell \geq \mu$. Expanding the determinant as

$$\begin{aligned}\det(I_3 - A_\alpha) &= (1-a_{11})(1-a_{22})(1-a_{33}) - a_{13}[a_{21}a_{32} + (1-a_{22})a_{31}]\\ &= (1-a_{22})\Big[(1-a_{11})(1-a_{33}) - a_{13}a_{31}\Big] - a_{13}a_{21}a_{32},\end{aligned}$$

we note that if the following is true for some $\Gamma > 1$,

$$-a_{13}a_{31} \geq -\frac{1}{\Gamma}(1-a_{11})(1-a_{33}), \tag{8}$$

$$-a_{13}a_{21}a_{32} \geq -\frac{\Gamma-1}{\Gamma(\Gamma+1)}(1-a_{11})(1-a_{22})(1-a_{33}), \tag{9}$$

then we obtain

$$\begin{aligned}\det(I_3 - A_\alpha) &\geq (1-a_{22})\Big[(1-a_{11})(1-a_{33}) - \frac{1}{\Gamma}(1-a_{11})(1-a_{33})\Big] - \frac{\Gamma-1}{\Gamma(\Gamma+1)}(1-a_{11})(1-a_{22})(1-a_{33})\\ &\geq (1-a_{22})(1-a_{11})(1-a_{33})\frac{\Gamma-1}{\Gamma} - \frac{\Gamma-1}{\Gamma(\Gamma+1)}(1-a_{11})(1-a_{22})(1-a_{33})\\ &\geq \left(\frac{\Gamma-1}{\Gamma+1}\right)(1-a_{22})(1-a_{11})(1-a_{33}) > 0,\end{aligned}$$

ensuring $\rho(A_\alpha) < 1$. We thus find the range of α that satisfies (8) and (9). Using $\{a_{ij}\}$'s from (6) in (8), we get

$$\begin{aligned}\alpha^2 q\left(g_3 + \alpha^2 g_4\right) &\leq \frac{1}{\Gamma}\left(\frac{1-\sigma_B^2}{2}\right)\left(\frac{1-\sigma_B^2}{6}\right)\\ \alpha^2 k_2(4 + \alpha^2 2\ell^2 y^2 \frac{2k_1 - 3k_2\alpha^2}{k_1 - 2k_2\alpha^2}(1+\beta\sigma_B)) &\leq \frac{1}{12\Gamma}\left(\frac{(1-\sigma_B^2)^3}{1+\sigma_B^2}\right)\\ \alpha^2 k_2\left(\frac{4k_1 - 8k_2\alpha^2) + 2\alpha^2\ell^2 y^2(1+\beta\sigma_B)(2k_1 - 3k_2\alpha^2)}{k_1 - 2k_2\alpha^2}\right) &\leq \frac{1}{12\Gamma}\left(\frac{(1-\sigma_B^2)^3}{1+\sigma_B^2}\right)\\ \alpha^2 k_2\left(4k_1 + 4k_1\alpha^2\ell^2 y^2(1+\beta\sigma_B)\right) + \frac{2k_2\alpha^2}{12\Gamma}\left(\frac{(1-\sigma_B^2)^3}{1+\sigma_B^2}\right) &\leq \frac{1}{36\Gamma}\left(\frac{(1-\sigma_B^2)^4}{1+\sigma_B^2}\right) + 8k_2^2\alpha^4 + 6k_2^2\alpha^6\ell^2 y^2(1+\beta\sigma_B)\\ \alpha^2 k_2\left(4k_1 + 4k_1\ell^2 y^2(1+\beta\sigma_B)\alpha^2 + \frac{2}{12\Gamma}\left(\frac{(1-\sigma_B^2)^3}{1+\sigma_B^2}\right)\right) &\leq \frac{1}{36\Gamma}\left(\frac{(1-\sigma_B^2)^4}{1+\sigma_B^2}\right) + 8k_2^2\alpha^4\\ &\quad + 6k_2^2\ell^2 y^2(1+\beta\sigma_B)\alpha^6\\ \alpha^2 k_1 k_2\left(4 + 4\ell^2 y^2(1+\beta\sigma_B)\alpha^2 + \frac{1}{2\Gamma}\left(\frac{(1-\sigma_B^2)^2}{1+\sigma_B^2}\right)\right) &\leq \frac{1}{36\Gamma}\left(\frac{(1-\sigma_B^2)^4}{1+\sigma_B^2}\right) + 8k_2^2\alpha^4\\ &\quad + 6k_2^2\ell^2 y^2(1+\beta\sigma_B)\alpha^6.\end{aligned}$$

We now simplify the above condition by letting $\alpha \leq \left(\frac{1-\sigma_B^2}{9\ell y_-}\right)\sqrt{\frac{\underline{\pi}}{\overline{\pi}}}$ in the LHS and decreasing the RHS, which leads to

$$\begin{aligned}
\alpha^2 &\leq \frac{\frac{1}{36\Gamma}\frac{(1-\sigma_B^2)^4}{1+\sigma_B^2}}{(2\ell^2 y_-^2\underline{\pi}^{-1}\overline{\pi}(1+\sigma_B^2))\left(4+4\left(\frac{(1-\sigma_B^2)\sqrt{\underline{\pi}}}{9y_-\sqrt{\overline{\pi}}}\right)^2 y^2(1+\beta\sigma_B)+\frac{1}{2\Gamma}\left(\frac{(1-\sigma_B^2)^2}{1+\sigma_B^2}\right)\right)} \\
&= \frac{\frac{(1-\sigma_B^2)^4}{1+\sigma_B^2}}{(\ell^2 y_-^2\underline{\pi}^{-1}\overline{\pi}(1+\sigma_B^2))\left(288\Gamma+288\Gamma\left(\frac{(1-\sigma_B^2)\sqrt{\underline{\pi}}}{9y_-\sqrt{\overline{\pi}}}\right)^2 y^2(1+\beta\sigma_B)+36\frac{(1-\sigma_B^2)^2}{1+\sigma_B^2}\right)} \\
\Longleftarrow \alpha^2 &\leq \frac{y_-^2\frac{(1-\sigma_B^2)^4}{1+\sigma_B^2}}{\ell^2 y_-^2 h(1+\sigma_B^2)\left(288y_-^2\Gamma+4\Gamma(1-\sigma_B^2)^2h^{-1}y^2(1+\beta\sigma_B)+36y_-^2\frac{(1-\sigma_B^2)^2}{1+\sigma_B^2}\right)} \\
&= \frac{y_-^2(1-\sigma_B^2)^4}{\ell^2 y_-^2 h(1+\sigma_B^2)\left(288y_-^2\Gamma(1+\sigma_B^2)+4\Gamma(1-\sigma_B^2)^2h^{-1}y^2(1+\beta\sigma_B)(1+\sigma_B^2)+36y_-^2(1-\sigma_B^2)^2\right)}.
\end{aligned}$$

We use $\sigma_B < 1, (1-\sigma_B^2) < 1, (1+\sigma_B^2) < 2, y^2(1+\beta) \geq 1, hy_-^2 \geq 1$ and $\Gamma hy^2(1+\beta) > 1$ leading to

$$\alpha^2 \leq \frac{y_-^2(1-\sigma_B^2)^4}{2\ell^2 y_-^2\left(612\Gamma hy_-^2y^2(1+\beta)+8\Gamma hy_-^2y^2(1+\beta)\right)} = \frac{(1-\sigma_B^2)^4}{1240\ell^2\left(\Gamma hy_-^2y^2(1+\beta)\right)}.$$

Taking square root of both sides results into

$$\alpha \leq \frac{(1-\sigma_B^2)^2}{36\ell y_- y\sqrt{\Gamma h(1+\beta)}}.$$

We next note that (9) holds when

$$\begin{aligned}
(\alpha^2 q)(\alpha^2 g_1+\alpha g_2)(\alpha^2 g_5) &\leq \frac{\Gamma-1}{\Gamma(\Gamma+1)}\left(1-\left(\frac{1+\sigma_B^2}{2}\right)\right)(1-(1-\alpha\mu))\left(1-\frac{5+\sigma_B^2}{6}\right) \\
\alpha^5 qg_5(\alpha g_1+g_2) &\leq \frac{\Gamma-1}{\Gamma(\Gamma+1)}\left(\frac{1-\sigma_B^2}{2}\right)(\alpha\mu)\left(\frac{1-\sigma_B^2}{6}\right) \\
\alpha^4 qg_5g_2(1+\alpha\mu) &\leq \frac{\Gamma-1}{\Gamma(\Gamma+1)}\left(\frac{1-\sigma_B^2}{2}\right)^2\left(\frac{\mu}{3}\right),
\end{aligned}$$

which can be simplified by using $\alpha \leq \frac{1}{\mu}$, i.e.,

$$\begin{aligned}
&\alpha^4 \leq \frac{\Gamma-1}{\Gamma(\Gamma+1)}\left(\frac{(1-\sigma_B^2)^3}{1+\sigma_B^2}\right)\left(\frac{\mu}{24}\right)\left(\frac{\mu}{\ell^6(18y_-^6y^2\underline{\pi}^{-1}\overline{\pi})(1+\beta\sigma_B)}\right) \\
\Longleftarrow\quad &\alpha^4 \leq \frac{\Gamma-1}{\Gamma(\Gamma+1)}\left(\frac{(1-\sigma_B^2)^3\mu^2}{864\ell^6(y_-^6y^2\underline{\pi}^{-1}\overline{\pi})(1+\beta\sigma_B)}\right) \\
\Longleftarrow\quad &\alpha \leq \frac{1}{6\ell\sqrt{\kappa}}\left[\frac{\Gamma-1}{\Gamma(\Gamma+1)}\left(\frac{(1-\sigma_B^2)^3}{y_-^6y^2h(1+\beta\sigma_B)}\right)\right]^{\frac{1}{4}},
\end{aligned}$$

for which it is sufficient to have

$$\alpha \leq \frac{(1-\sigma_B^2)^{3/4}}{12\ell\sqrt{\kappa}}\left(\frac{\Gamma-1}{\Gamma^2y_-^6y^2h(1+\beta)}\right)^{\frac{1}{4}}. \tag{10}$$

We next select the minimum of all the bounds on step-size,

$$\begin{aligned}
&\alpha \le \min\left\{\frac{1-\sigma_B^2}{9\ell y_- \sqrt{h}}, \frac{(1-\sigma_B^2)^2}{36\ell y_- y\sqrt{\Gamma h(1+\beta)}}, \frac{(1-\sigma_B^2)^{3/4}}{12\ell\sqrt{\kappa}}\left(\frac{\Gamma-1}{\Gamma^2 y_-^6 y^2 h(1+\beta)}\right)^{\frac{1}{4}}\right\} \\
\Longleftarrow\ & \alpha \le \frac{(1-\sigma_B^2)^2}{36\ell\sqrt{\kappa}}\cdot\min\left\{\left(\frac{1}{\tau\Gamma}\right)^{\frac{1}{2}}, \left(\frac{\Gamma-1}{\tau\Gamma^2}\right)^{\frac{1}{4}}\right\} \\
\Longleftarrow\ & \alpha \le \frac{(1-\sigma_B^2)^2}{36\ell\sqrt{\kappa}}\cdot\frac{1}{\sqrt{\tau\Gamma}}\cdot\min\left\{1, (\Gamma-1)^{\frac{1}{4}}\right\},
\end{aligned}$$

where $\tau := y_-^6 y^2 h(1+\beta)$. We note that the above is true for all $\Gamma > 1$ and $\min\left\{1, (\Gamma-1)^{\frac{1}{4}}\right\}$ is maximized at $\Gamma = 2$. Hence, for a largest possible α, that is feasible given our bound, we select $\Gamma = 2$, which leads to

$$\alpha \le \frac{1}{\ell\sqrt{\kappa}}\cdot\frac{(1-\sigma_B^2)^2}{51\sqrt{\tau}}.$$

Thus, when α follows the above relation, we have $\rho(A_\alpha) < 1$ and using the linear system recursion in (7), we get

$$\lim_{k\to\infty} \mathbf{t}_{k+1} \le (I_3 - A_\alpha)^{-1}\mathbf{c}, \tag{11}$$

since $\lim_{k\to\infty} H_k$ is a zero matrix. The first two elements in the R.H.S (vector) of (11) can be manipulated as follows:

$$\begin{aligned}
[(I_3 - A_\alpha)^{-1}\mathbf{c}]_1 &= \frac{a_{13}a_{32}\frac{\alpha^2\sigma^2}{n} + a_{13}(1-a_{22})C_\sigma}{\det(I_3 - A_\alpha)} \\
&\le \left(\frac{\Gamma+1}{\Gamma-1}\right)\frac{a_{13}}{(1-a_{11})(1-a_{22})(1-a_{33})}\left[a_{32}\frac{\alpha^2\sigma^2}{n} + (1-a_{22})C_\sigma\right] \\
&\le \left(\frac{\alpha^2\left(\frac{1+\sigma_B^2}{1-\sigma_B^2}\right)}{\left(\frac{1-\sigma_B^2}{2}\right)(\alpha\mu)\left(\frac{1-\sigma_B^2}{6}\right)}\right)\left[\alpha^2(18\ell^4 y_-^4 y^2 \underline{\pi}^{-1})\left(\frac{1+\sigma_B^2}{1-\sigma_B^2}\right)\left(\frac{\alpha^2\sigma^2}{n}\right) + (\alpha\mu)C_\sigma\right] \\
&\le \left(\frac{12\alpha\left(1+\sigma_B^2\right)}{\mu\left(1-\sigma_B^2\right)^3}\right)\left[18\alpha^4\ell^4 y_-^4 y^2\underline{\pi}^{-1}\left(\frac{1+\sigma_B^2}{1-\sigma_B^2}\right)\left(\frac{\sigma^2}{n}\right) + \alpha\mu(4\sigma^2 n\underline{\pi}^{-1})\left(\frac{1+\sigma_B^2}{1-\sigma_B^2}\right)\right] \\
&= \alpha^5\left(\frac{\ell^4\sigma^2}{n\mu}\right)\left(\frac{216 y_-^4 y^2\underline{\pi}^{-1}\left(1+\sigma_B^2\right)^2}{\left(1-\sigma_B^2\right)^4}\right) + \alpha^2(n\sigma^2)\left(\frac{48\underline{\pi}^{-1}\left(1+\sigma_B^2\right)^2}{\left(1-\sigma_B^2\right)^4}\right) \\
&= \frac{\alpha^5}{(1-\sigma_B^2)^4}\mathcal{O}\left(\frac{\ell^4\sigma^2}{n\mu}\right) + \frac{\alpha^2}{(1-\sigma_B^2)^4}\mathcal{O}\left(n\sigma^2\right);
\end{aligned}\tag{12}$$

$$\begin{aligned}
[(I_3 - A_\alpha)^{-1}\mathbf{c}]_2 &= \frac{[(1-a_{11})(1-a_{33}) - a_{13}a_{31}]\frac{\alpha^2\sigma^2}{n} + (a_{13}a_{21})C_\sigma}{\det(I_3 - A_\alpha)} \\
&\le \frac{\Gamma+1}{\Gamma}\left(\frac{\alpha^2\sigma^2}{n(1-a_{22})}\right) + \left(\frac{\Gamma+1}{\Gamma-1}\right)\left(\frac{a_{13}a_{21}C_\sigma}{(1-a_{11})(1-a_{22})(1-a_{33})}\right) \\
&\le \frac{\alpha^2\sigma^2}{n(\alpha\mu)} + \frac{\alpha^2\left(\frac{1+\sigma_B^2}{1-\sigma_B^2}\right)\left(\alpha^2\left(\frac{\ell^2 y_-^2(1+\beta\sigma_B)\overline{\pi}}{n}\right) + \alpha\left(\frac{\ell^2 y_-^2(1+\beta\sigma_B)\overline{\pi}}{n\mu}\right)\right)C_\sigma}{\left(\frac{1-\sigma_B^2}{2}\right)(\alpha\mu)\left(\frac{1-\sigma_B^2}{6}\right)} \\
&= \frac{\alpha\sigma^2}{n\mu} + \frac{12\alpha\left(1+\sigma_B^2\right)^2\left(\alpha^2(\ell^2 y_-^2(1+\beta\sigma_B)\overline{\pi}) + \alpha\left(\frac{\ell^2 y_-^2(1+\beta\sigma_B)\overline{\pi}}{\mu}\right)\right)(4\sigma^2 n\underline{\pi}^{-1})}{n\mu(1-\sigma_B^2)^4} \\
&= \alpha\mathcal{O}\left(\frac{\sigma^2}{n\mu}\right) + \frac{\alpha^2}{(1-\sigma_B^2)^4}\mathcal{O}\left(\frac{\ell^2\sigma^2}{\mu^2}\right).
\end{aligned}\tag{13}$$

Finally, the mean network error, defined as $\mathbf{e}_k := \frac{1}{n}\mathbb{E}\left[\|\mathbf{z}_k - \mathbf{1}_n\mathbf{z}^*\|_2^2\right]$, is given by

$$\mathbf{e}_k \leq \frac{3y_-^2\overline{\pi}}{n}\mathbb{E}[\|\mathbf{x}_k - B^\infty\mathbf{x}_k\|_{\boldsymbol{\pi}}^2] + 3y_-^2\beta^2\mathbb{E}[\|\mathbf{z}^*\|_2^2]\sigma_B^{2k} + 3y_-^2y^2\mathbb{E}[\|\mathbf{x}_k - \mathbf{1}_n\mathbf{z}^*\|_2^2]. \tag{14}$$

Notice that the second term of (14) vanishes asymptotically. Using (12) and (13), we further have

$$\limsup_{k\to\infty}\mathbf{e}_k \leq \frac{3y_-^2\overline{\pi}\alpha^5}{(1-\sigma_B^2)^4}\mathcal{O}\left(\frac{\ell^4\sigma^2}{n^2\mu}\right) + \frac{3y_-^2\overline{\pi}\alpha^2}{(1-\sigma_B^2)^4}\mathcal{O}\left(\sigma^2\right) + \frac{3y_-^2y^2\alpha^2}{(1-\sigma_B^2)^4}\mathcal{O}\left(\frac{\ell^2\sigma^2}{\mu^2}\right) + 3y_-^2y^2\alpha\mathcal{O}\left(\frac{\sigma^2}{n\mu}\right). \tag{15}$$

and the theorem follows by dropping the higher order term of α and noting that $\frac{\ell^2}{\mu^2} \geq 1$. □

Corollary 2. *For all $k, \exists b \in \mathbb{R}$, such that $\mathbb{E}[\|\mathbf{x}_k\|_2^2] \leq b$.*

The proof follows from Theorem 1.

B. Proof of Theorem 2

Let $P_k := \mathbb{E}[\|\mathbf{x}_k - B^\infty\mathbf{x}_k\|_{\boldsymbol{\pi}}^2]$, $Q_k := \mathbb{E}[\|\overline{\mathbf{x}}_k - \mathbf{z}^*\|_2^2]$ and $R_k := \mathbb{E}[\|\mathbf{w}_k - B^\infty\mathbf{w}_k\|_{\boldsymbol{\pi}}^2]$. To show that

$$P_k \leq \frac{\widetilde{P}}{(m+k)^2}, \qquad Q_k \leq \frac{\widetilde{Q}}{(m+k)}, \qquad R_k \leq \widetilde{R}, \tag{16}$$

for all $k \geq 0$, it suffices to show that the R.H.S of (7), with a decaying step-size $\alpha_k < \left(\frac{1-\sigma_B^2}{6\ell}\right)\frac{1}{y_-\sqrt{(1+\sigma_B^2)h}}$, follows the above bounds. We develop the proof by induction. Consider (7) for $k=0$, i.e.,

$$A_{\alpha_0}\mathbf{t}_0 + H_0\mathbf{s}_0 + \mathbf{c}$$

with $\alpha_0 = \frac{\theta}{m}$, and therefore $m > \frac{6\ell\theta y_-\sqrt{(1+\sigma_B^2)h}}{1-\sigma_B^2}$, to obtain the following conditions:

$$\widetilde{R} \leq \left(\frac{1-\sigma_B^2}{\theta^2(1+\sigma_B^2)}\right)\left(\frac{m^2}{(m+1)^2} - \frac{1+\sigma_B^2}{2}\right)\widetilde{P}, \tag{17a}$$

$$\widetilde{Q} \geq \left[\left(\frac{\theta}{m} + \frac{1}{\mu}\right)\left(\frac{\theta\ell^2E_1}{mn\left(\theta\mu-1\right)}\right)\widetilde{P} + \frac{nm^2K_1b + \theta^2\sigma^2}{n\left(\theta\mu-1\right)}\right], \tag{17b}$$

$$\widetilde{R} \geq \frac{6}{1-\sigma_B^2}\left[\left(\frac{E_2}{m^2}\right)\widetilde{P} + \left(\frac{\theta^2\ell^4E_3}{m^3}\right)\widetilde{Q} + K_2b + C_0\right]. \tag{17c}$$

where E_1, E_2, E_3 are defined in the following. It can be verified, that the above conditions hold if and only if

$$\frac{1-\sigma_B^2}{\theta^2(1+\sigma_B^2)}\left(\frac{m^2}{(m+1)^2} - \frac{1+\sigma_B^2}{2}\right)\widetilde{P} > \frac{6}{1-\sigma_B^2}\left[\frac{E_2\widetilde{P}}{m^2} + \left(\frac{\theta^2\ell^4E_3}{m^3}\right)\left(\frac{\theta}{m} + \frac{1}{\mu}\right)\left(\frac{\theta\ell^2E_1\widetilde{P}}{mn\left(\theta\mu-1\right)}\right)\right]$$

$$\frac{(1-\sigma_B^2)^2}{6\theta^2(1+\sigma_B^2)}\left(\frac{1-\sigma_B^2}{2} - \frac{2m+1}{(m+1)^2}\right) > \frac{E_2}{m^2} + \left(\frac{\theta^3\ell^6E_1E_3}{m^4n\left(\theta\mu-1\right)}\right)\left(\frac{\theta}{m} + \frac{1}{\mu}\right)$$

and $\widetilde{Q} = \max\{mQ_0, D_6\}$, where $\widetilde{P}$ and $\widetilde{R}$ follow the constraints in (17a), (17c), and $\widetilde{R} > R_0$. We use $\mathbb{E}\|\mathbf{x}_k\|_2^2 < b$, for some $b > 0$, which follows from Theorem 1. Thus, $\widetilde{P}$ is selected as $\widetilde{P} = \max\left\{m^2 P_0, \frac{R_0}{D_1}, \frac{D_3}{D_1 - D_2}, \frac{D_5}{D_1 - D_4}\right\}$, where

$$C_0 := 4\sigma^2 q \underline{\pi}^{-1}\left(n + 3\left(\frac{\theta^2\ell^2 y_-^4 y^2}{m^2}\right)\left(\frac{2m^2k_1 - 3k_2\theta^2}{m^2k_1 - 2k_2\theta^2}\right)\right),$$

$$D_1 := \left(\frac{1-\sigma_B^2}{\theta^2(1+\sigma_B^2)}\right)\left(\frac{1-\sigma_B^2}{2} - \frac{2m+1}{(m+1)^2}\right),$$

$$D_3 := \left[\frac{6}{1-\sigma_B^2}\right]\left[\left(\frac{\theta^2\ell^4 E_3}{m^3}\right)\|\overline{\mathbf{x}}_0 - \mathbf{z}^*\|_2^2 + C_0 + K_2 b\right],$$

$$D_5 := \left[\frac{6}{1-\sigma_B^2}\right]\left[\left(\frac{\theta^2\ell^4 E_3}{m^3 n(\theta\mu - 1)}\right)(\theta^2\sigma^2 + nm^2K_1 b) + C_0 + K_2 b\right],$$

$$D_6 := \left[\frac{1}{n(\theta\mu - 1)}\right]\left[\left(\frac{\theta}{m} + \frac{1}{\mu}\right)\left(\frac{\theta\ell^2 E_1}{m}\right)\widetilde{P} + \theta^2\sigma^2 + nm^2K_1 b\right],$$

$$K_2 := \frac{12\ell^2 q y_-^4 \beta}{\overline{\pi}}\left(2\beta + \frac{\theta^2\ell^2 y_-^2 y^2(\beta+1)}{m^2}\left(\frac{2m^2k_1 - 3k_2\theta^2}{m^2k_1 - 2k_2\theta^2}\right)\right),$$

$$E_2 := 4k_2 + \left(\frac{2\ell^2 y^2 k_2\theta^2}{m^2}\right)\left(\frac{2k_1m^2 - 3k_2\theta^2}{k_1m^2 - 2k_2\theta^2}\right)(1 + \beta\sigma_B).$$

$$D_2 := \frac{6E_2}{m^2(1-\sigma_B^2)},$$

$$D_4 := \left[\frac{6E_1}{1-\sigma_B^2}\right]\left(\frac{\theta^3\ell^6 E_3}{m^4 n(\theta\mu - 1)}\right)\left(\frac{\theta}{m} + \frac{1}{\mu}\right) + D_2,$$

$$E_1 := (1 + \beta\sigma_B)y_-^2\overline{\pi},$$

$$E_3 := 18 q y_-^4 y^2 \underline{\pi}^{-1},$$

$$K_1 := K_3\left(\frac{\theta\ell^2}{\mu m} + \frac{\theta^2\ell^2}{m^2}\right),$$

$$K_3 := y_-^2\beta(\beta + 1),$$

Thus, we conclude that (16) holds for $k = 0$ when the corresponding conditions on $\widetilde{P}, \widetilde{Q}, \widetilde{R}$, and m are met. Next, assume that (16) holds for some k, it can be verified that it automatically holds for $k+1$ with the same conditions on $\widetilde{P}, \widetilde{Q}, \widetilde{R}$, and m that are derived for $k = 0$.

We next improve Q_k to establish the network-independence. Pick $\widetilde{S}$ large enough such that $\forall k \geq \widetilde{S}, \sigma_B^k \leq \frac{1}{n(m+k)^2}$. Then using the decaying step-size in (7), we have

$$Q_{k+1} \leq \left(1 - \frac{\theta\mu}{m+k}\right)Q_k + \frac{2\theta\ell^2(E_1\widetilde{P} + K_3 b)}{n\mu(m+k)^3} + \frac{\theta^2\sigma^2}{n(m+k)^2},$$

which leads to

$$Q_k \leq \prod_{t=0}^{k-1}\left(1 - \frac{\theta\mu}{m+t}\right)Q_0 + \sum_{t=0}^{k-1}\prod_{j=t+1}^{k-1}\frac{m+j-\theta}{m+j}\left[\frac{2\theta\ell^2(E_1\widetilde{P} + K_3 b)}{n\mu(m+t)^3} + \frac{\theta^2\sigma^2}{n(m+t)^2}\right], \tag{18}$$

From [17], we have

$$\prod_{t=0}^{k-1}\left(1 - \frac{\theta\mu}{m+t}\right) \leq \frac{m^{\theta\mu}}{(m+k)^{\theta\mu}}, \qquad \prod_{j=t+1}^{k-1}\left(1 - \frac{\theta\mu}{m+j}\right) \leq \frac{(m+t+1)^{\theta\mu}}{(m+k)^{\theta\mu}};$$

Using the above relations and in (18),

$$\begin{aligned}
Q_k &\leq \frac{m^{\theta\mu}}{(m+k)^{\theta\mu}}Q_0 + \frac{4\theta\ell^2(E_1\widetilde{P} + K_3 b)}{n\mu(m+k)^{\theta\mu}}\sum_{t=0}^{k-1}(m+t)^{\theta\mu-3} + \frac{2\theta^2\sigma^2}{n(m+k)^{\theta\mu}}\sum_{t=0}^{k-1}(m+t)^{\theta\mu-2} \\
&\leq \frac{m^{\theta\mu}}{(m+k)^{\theta\mu}}Q_0 + \frac{4\theta\ell^2(E_1\widetilde{P} + K_3 b)}{n\mu(m+k)^{\theta\mu}}\int_{t=-1}^{k}(m+t)^{\theta\mu-3}dt + \frac{2\theta^2\sigma^2}{n(m+k)^{\theta\mu}}\int_{t=-1}^{k}(m+t)^{\theta\mu-2}dt \\
&\leq \frac{2\theta^2\sigma^2}{n(\theta\mu-1)(m+k)} + \frac{m^{\theta\mu}}{(m+k)^{\theta\mu}}Q_0 + \max\left\{\frac{4\theta\ell^2(E_1\widetilde{P} + K_3 b)}{n\mu(\theta\mu-2)(m+k)^2}, \frac{4\theta\ell^2(E_1\widetilde{P} + K_3 b)(m-1)^{\theta\mu-2}}{n\mu(2-\theta\mu)\mu(m+k)^{\theta\mu}}\right\},
\end{aligned}$$

and the theorem follows by (3) in Lemma 2 and by noting that the $\frac{1}{(m+k)}$ term in Q_k is network independent. $\square$

V. Numerical Simulations

In this section, we illustrate **S-ADDOPT** and compare its performance with related algorithms over directed graphs, i.e., **GP** [20], [21], **ADDOPT** [14], [31], and **SGP** [16], [25], [26]. Recall that **GP** and **ADDOPT** are batch algorithms and operate on the entire local batch of data at each node. In other words, the true gradient ∇f_i is used at each node to compute the algorithm updates. In contrast, **SGP** and **S-ADDOPT** employ a stochastic gradient $\nabla \widehat{f}_i(\cdot) = \nabla f_{i,s_k^i}(\cdot)$, where s_k^i is chosen uniformly at random from the index set $\{1,\ldots,m_i\}$ at each node i and each time k. It can be verified that this choice of stochastic gradient satisfies the $\texttt{SFO}$ setup in Assumption 3. The numerical experiments are described next.

A. *Logistic Regression: Strongly convex*

We now show the numerical experiments for a binary classification problem to classify hand-written digits $\{3,8\}$ from the MNIST dataset. In this setup, there are a total of $N = 12{,}000$ labeled images for training and each node i possesses a local batch with m_i training samples. The j-th sample at node i is a tuple $\{\mathbf{x}_{i,j}, y_{i,j}\} \subseteq \mathbb{R}^{784} \times \{+1,-1\}$ and the local logistic regression cost function f_i at node i is given by

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

which is smooth and strongly convex because of the addition of the regularizer λ. The nodes cooperate to solve the following decentralized optimization problem:

$$\min_{\mathbf{b}\in\mathbb{R}^{784},\, c\in\mathbb{R}} F(\mathbf{b},c) = \frac{1}{n}\sum_i f_i.$$

For all algorithms, the step-sizes are hand-tuned for best performance. The column stochastic weights are chosen such that $b_{ji} = |\mathcal{N}_i^{\text{out}}|^{-1}$, for each $j \in \mathcal{N}_i^{\text{out}}$.

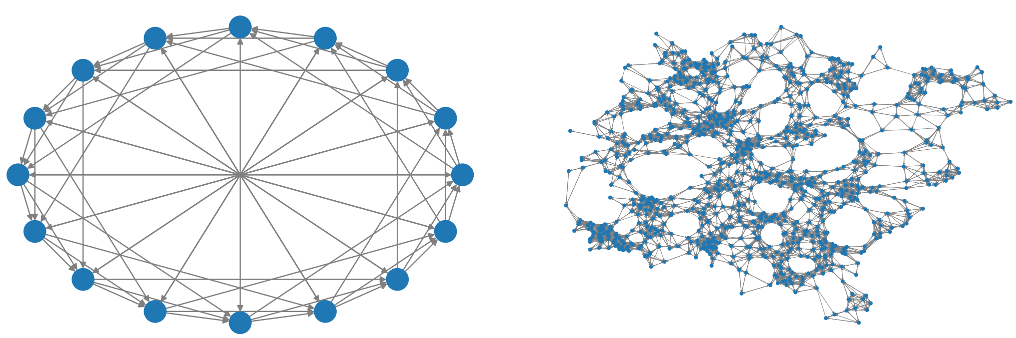

Fig. 1. (Left) Directed exponential graph with $n = 16$ nodes. (Right) geometric graph with $n = 1000$ nodes

Structured training setup–Data-centers: We choose an exponential graph with $n = 16$ nodes (Fig. 1, left) to model a highly structured communication graph mimicking, for example, a data center where the data is typically evenly divided among the nodes. In particular, we choose $m_i = N/n = 750$ training images at each node i. Performance comparison is provided in Fig. 2, for a constant step-size, and in Fig. 4 (left), for decaying step-sizes, where we plot the optimality gap $F(\overline{\mathbf{x}}_k) - F(\mathbf{z}^*)$ versus the number of epochs. Each epoch represents $N/n = 750$ stochastic gradient evaluations implemented (in parallel) at each node. Recall that **S-ADDOPT** adds gradient tracking to **SGP** and in this balanced data scenario, its performance is virtually indistinguishable from **SGP**, while their batch counterparts are much slower. **ADDOPT** however converges linearly to the exact solution as can be observed in Fig. 2 (right) over a longer number of epochs.

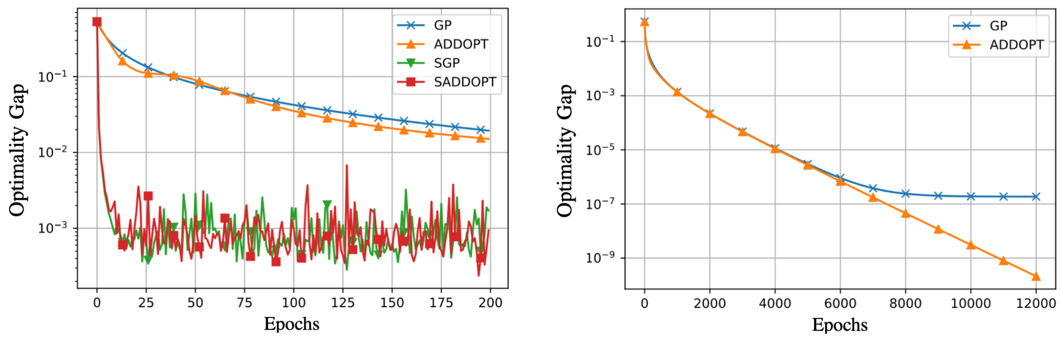

Fig. 2. (Left) Balanced data and constant step-sizes for all algorithms: Performance comparison over the exponential graph with $n = 16$ nodes and $m = 750$ data samples per node. (Right) Linear convergence of **ADDOPT** shown over a longer number of epochs.

Ad hoc training setup–Multi-agent networks: We next consider a large-scale nearest-neighbor (geometric) digraph with $n = 1{,}000$ nodes (Fig. 1, right) that models, for example, ad hoc wireless multi-agent networks, where the agents typically possess different sizes of local batches depending on their locations and local resources; see Fig. 3 (left) for an arbitrary data distribution across the agents. Performance comparison is shown in Fig. 3 (right), for a constant step-size, and in Fig. 4 (right), for decaying step-sizes. Each epoch represents $N/n = 12$ component gradient evaluations (in parallel) at each node. When the data is unbalanced, the addition of gradient tracking in **S-ADDOPT** results in a significantly improved performance than **SGP**.

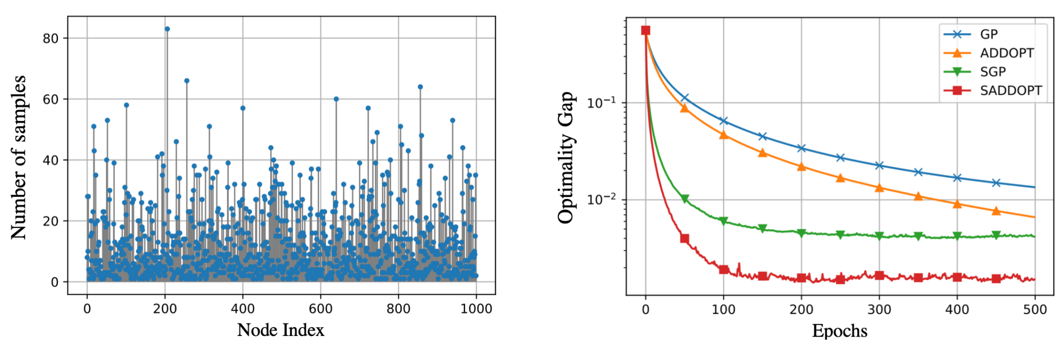

Fig. 3. Performance comparison (right), over the directed geometric graph in Fig. 1 (right), with an unbalanced data distribution (left) and constant step-sizes for all algorithms.

Comparing the structured and ad hoc training scenarios, we note that gradient tracking does not show a noticeable improvement over the balanced data scenario but results in a superior performance when the data distribution is unbalanced. This is because the convergence (15) of **S-ADDOPT** (similar to its undirected counterpart [12]) does not depend on the heterogeneity of local data batches as opposed to **SGP**. A detailed discussion along these lines can be found in [19].

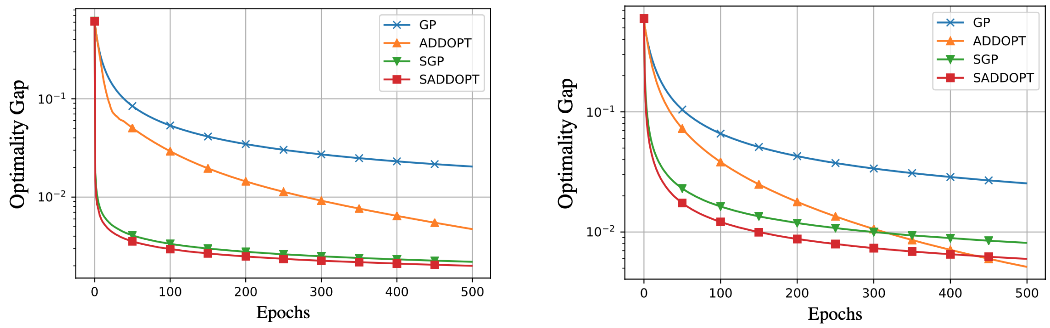

Fig. 4. Performance comparison for exact convergence (decaying step-sizes for **S-ADDOPT** and **SGP**, and constant step-size for **ADDOPT**): (Left) Directed exponential graph with balanced data. (Right) Directed geometric graph with unbalanced data.

B. Neural networks: Non-convex

Finally, we compare the performance of the stochastic algorithms discussed in this report for training a distributed neural network optimizing a non-convex problem with constant step-sizes of the algorithms. Each node has a local neural network comprising of one fully connected hidden layer of 64 neurons learning 51,675 parameters. We train the neural network to for a multi-class classification problem to classify ten classes in MNIST $\{0, \cdots, 9\}$ and CIFAR-10 {“airplanes”, $\cdots$, “trucks”} datasets. Both have 60,000 images in total and 6,000 images per class. The data samples are divided randomly and equally over a 500 node directed geometric graph shown in Fig. 5. We

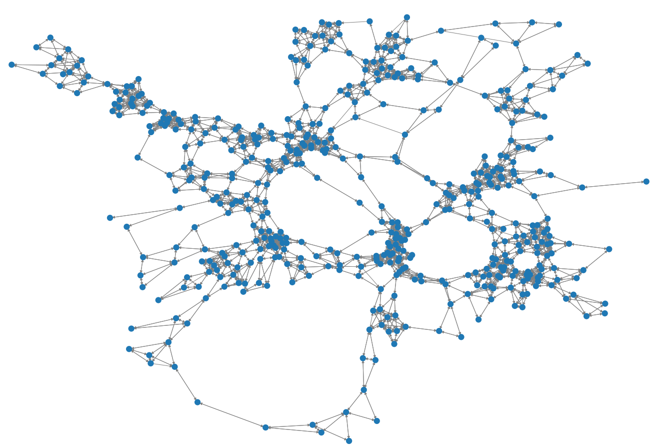

Fig. 5. Directed geometric graph with $n = 500$ nodes.

show the loss $F(\overline{\mathbf{x}}_k)$ and test accuracy of **SGP** and **S-ADDOPT** with respect to epochs over the MNIST dataset in Fig. 6. Similarly, Fig. 7 illustrates the performance for the CIFAR-10 dataset. We observe that adding gradient tracking in **SGP** improves the transient and steady state performance in these non-convex problems.

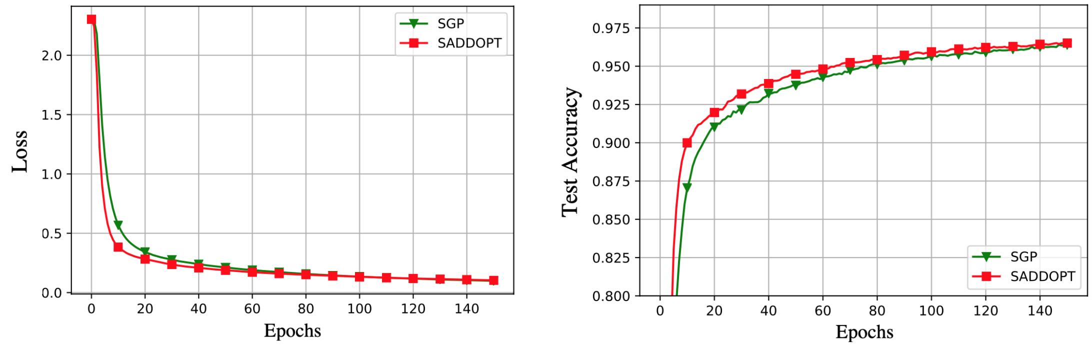

Fig. 6. MNIST classification using a two-layer neural network over a directed geometric graph with $n = 500$ nodes and $m = 120$ data samples per node; both algorithms use a constant step-size.

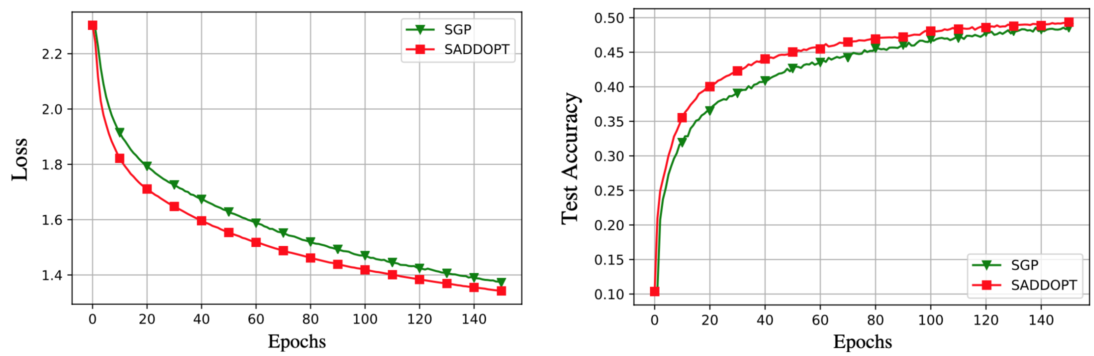

Fig. 7. CIFAR-10 classification using a two-layer neural network over a directed geometric graph with $n = 500$ nodes and $m = 120$ data samples per node; both algorithms use a constant step-size.

VI. Conclusions

In this report, we present **`S-ADDOPT`**, a decentralized stochastic optimization algorithm that is applicable to both undirected and directed graphs. **`S-ADDOPT`** adds gradient tracking to **`SGP`** and can be viewed as a stochastic extension of **`ADDOPT`**. We show that for a constant step-size α, **`S-ADDOPT`** converges linearly inside an error ball around the optimal, the size of which is controlled by α. For decaying step-sizes $\mathcal{O}(1/k)$, we show that **`S-ADDOPT`** is asymptotically network-independent and reaches the exact solution sublinearly at $\mathcal{O}(1/k)$. These characteristics match the centralized **`SGD`** up to some constant factors. Numerical experiments over both strongly convex and non-convex problems illustrate the convergence behavior and the performance comparison of **`S-ADDOPT`** versus **`SGP`** and their non-stochastic counterparts.

References

[1] S. Lee and M. M. Zavlanos, "Approximate projection methods for decentralized optimization with functional constraints," *IEEE Transactions on Automatic Control*, vol. 63, no. 10, pp. 3248–3260, Oct. 2018.

[2] S. Safavi, U. A. Khan, S. Kar, and J. M. F. Moura, "Distributed localization: A linear theory," *Proceedings of the IEEE*, vol. 106, no. 7, pp. 1204–1223, Jul. 2018.

[3] P. A. Forero, A. Cano, and G. B. Giannakis, "Consensus-based distributed support vector machines," *Journal of Machine Learning Research*, vol. 11, pp. 1663–1707, 2010.

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

[5] T. Yang, X. Yi, J. Wu, Y. Yuan, D. Wu, Z. Meng, Y. Hong, H. Wang, Z. Lin, and K. H. Johansson, "A survey of distributed optimization," *Annual Reviews in Control*, vol. 47, pp. 278 – 305, 2019.

[6] 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.

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

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

[9] 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.

[10] 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.

[11] R. Xin, U. A. Khan, and S. Kar, "Variance-reduced decentralized stochastic optimization with accelerated convergence," *arXiv:1912.04230*, 2019.

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

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

[14] 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.

[15] 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. 315–320, 2018.

[16] A. Spiridonoff, A. Olshevsky, and I. C. Paschalidis, "Robust asynchronous stochastic gradient-push: Asymptotically optimal and network-independent performance for strongly convex functions," *Journal of Machine Learning Research*, vol. 21, no. 58, pp. 1–47, 2020.

[17] S. Pu, A. Olshevsky, and I. C. Paschalidis, "A sharp estimate on the transient time of distributed stochastic gradient descent," *1906.02702*, 2019.

[18] S. Pu, A. Olshevsky, and I. C. Paschalidis, "Asymptotic network independence in distributed stochastic optimization for machine learning: Examining distributed and centralized stochastic gradient descent," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 114–122, 2020.

[19] 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.

[20] 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.

[21] A. Nedić and A. Olshevsky, "Distributed optimization over time-varying directed graphs," *IEEE Transactions on Automatic Control*, vol. 60, no. 3, pp. 601–615, 2014.

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

[23] 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 Transactions on Automatic Control*, vol. 63, no. 10, pp. 3558–3565, Oct. 2018.

[24] R. Xin, C. Xi, and U. A. Khan, "FROST—Fast row-stochastic optimization with uncoordinated step-sizes," *EURASIP Journal on Advances in Signal Processing*, Jan. 2019.

[25] 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.

[26] M. Assran, N. Loizou, N. Ballas, and M. Rabbat, "Stochastic gradient push for distributed deep learning," in *36th International Conference on Machine Learning*. Jun. 2019, vol. 97, pp. 344–353, PMLR.

[27] 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 on Decision and Control*, Dec. 2019, pp. 8353–8358.

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

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

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

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

[32] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, 1985.

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

[34] 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.

APPENDIX A
DEVELOPING THE LTI SYSTEM DESCRIBING S-ADDOPT

To derive the LTI system described in (7), we first define a few terms:

$$\overline{\mathbf{w}}_k := \frac{1}{n}\mathbf{1}_n^\top \mathbf{w}_k, \quad \overline{\mathbf{h}}_k := \frac{1}{n}\mathbf{1}_n^\top \nabla f(\mathbf{z}_k), \quad \overline{\mathbf{g}}_k := \frac{1}{n}\mathbf{1}_n^\top \nabla \widehat{f}(\mathbf{z}_k) := \overline{\mathbf{w}}_k,$$
$$\overline{\mathbf{p}}_k := \frac{1}{n}\mathbf{1}_n^\top \nabla f(\mathbf{1}_n \overline{\mathbf{x}}_k), \quad \nabla f(\mathbf{z}_k) := [\nabla f_1(\mathbf{z}_k^1)^\top, \cdots, \nabla f_n(\mathbf{z}_k^n)^\top]^\top.$$

We denote $\xi_k^i \in \mathbb{R}^p$ as random vectors for all $k \geq 0$ and $i \in \mathcal{V}$ such that the stochastic gradient is $\nabla \widehat{f}_i(\mathbf{z}_k^i) = \nabla f_i(\mathbf{z}_k^i, \xi_k^i)$. Assumption 3 allows the gradient noise processes to be dependent on agent i and the current iterate $\mathbf{z}_k^i$. We denote by $\mathcal{F}_k$, the σ-algebra generated by the set of random vectors $\{\xi_l^i\}_{i\in\mathcal{V}}$, where $0 \leq l \leq k-1$. The derivation of the system described in (7) is now provided in the following three steps:

Step 1. Network agreement error.

Note that the first term $\|\mathbf{x}_{k+1} - B^\infty \mathbf{x}_{k+1}\|_{\boldsymbol{\pi}}^2$ in the LTI system is essentially the network agreement error and it can be expanded as:

$$\begin{aligned}
\|\mathbf{x}_{k+1} - B^\infty \mathbf{x}_{k+1}\|_{\boldsymbol{\pi}}^2 &= \|B\mathbf{x}_k - B^\infty \mathbf{x}_k - \alpha(\mathbf{w}_k - B^\infty \mathbf{w}_k)\|_{\boldsymbol{\pi}}^2 \\
&= \|B\mathbf{x}_k - B^\infty \mathbf{x}_k\|_{\boldsymbol{\pi}}^2 + \alpha^2\|\mathbf{w}_k - B^\infty \mathbf{w}_k\|_{\boldsymbol{\pi}}^2 - 2\langle B\mathbf{x}_k - B^\infty \mathbf{x}_k, \alpha(\mathbf{w}_k - B^\infty \mathbf{w}_k)\rangle_{\boldsymbol{\pi}} \\
&\leq \sigma_B^2\|\mathbf{x}_k - B^\infty \mathbf{x}_k\|_{\boldsymbol{\pi}}^2 + \alpha^2\|\mathbf{w}_k - B^\infty \mathbf{w}_k\|_{\boldsymbol{\pi}}^2 + 2\alpha\sigma_B\|\mathbf{x}_k - B^\infty \mathbf{x}_k\|_{\boldsymbol{\pi}}\|\mathbf{w}_k - B^\infty \mathbf{w}_k\|_{\boldsymbol{\pi}} \\
&\leq \left(\sigma_B^2 + \alpha\sigma_B\frac{1-\sigma_B^2}{2\alpha\sigma_B}\right)\|\mathbf{x}_k - B^\infty \mathbf{x}_k\|_{\boldsymbol{\pi}}^2 + \left(\alpha^2 + \alpha\sigma_B\frac{2\alpha\sigma_B}{1-\sigma_B^2}\right)\|\mathbf{w}_k - B^\infty \mathbf{w}_k\|_{\boldsymbol{\pi}}^2 \\
&= \left(\frac{1+\sigma_B^2}{2}\right)\|\mathbf{x}_k - B^\infty \mathbf{x}_k\|_{\boldsymbol{\pi}}^2 + \alpha^2\left(\frac{1+\sigma_B^2}{1-\sigma_B^2}\right)\|\mathbf{w}_k - B^\infty \mathbf{w}_k\|_{\boldsymbol{\pi}}^2.
\end{aligned} \tag{19}$$

Step 2. Optimality gap.

Next, we consider $\|\overline{\mathbf{x}}_{k+1} - \mathbf{z}^*\|_2^2$, which defines the the gap between the mean iterate and the true solution:

$$\|\overline{\mathbf{x}}_{k+1} - \mathbf{z}^*\|_2^2 = \|(\overline{\mathbf{x}}_k - \alpha\overline{\mathbf{w}}_k) - \mathbf{z}^*\|_2^2 = \|\overline{\mathbf{x}}_k - \mathbf{z}^*\|_2^2 + \alpha^2\|\overline{\mathbf{g}}_k\|_2^2 - 2\langle \overline{\mathbf{x}}_k - \mathbf{z}^*, \overline{\mathbf{g}}_k\rangle.$$

Noticing that $\mathbb{E}[\overline{\mathbf{g}}_k|\mathcal{F}_k] = \overline{\mathbf{h}}_k$,

$$\mathbb{E}[\|\overline{\mathbf{g}}_k\|_2^2|\mathcal{F}_k] = \mathbb{E}[\|\overline{\mathbf{g}}_k - \overline{\mathbf{h}}_k\|_2^2|\mathcal{F}_k] + \|\overline{\mathbf{h}}_k\|_2^2 \leq \frac{\sigma^2}{n} + \|\overline{\mathbf{h}}_k\|_2^2.$$

For $\eta = (1-\alpha\mu)$, we can write:

$$\begin{aligned}
\mathbb{E}[\|\overline{\mathbf{x}}_{k+1} - \mathbf{z}^*\|_2^2|\mathcal{F}_k] &\leq \|\overline{\mathbf{x}}_k - \mathbf{z}^*\|_2^2 - 2\langle \overline{\mathbf{x}}_k - \mathbf{z}^*, \overline{\mathbf{h}}_k\rangle + \alpha^2\|\overline{\mathbf{h}}_k\|_2^2 + \frac{\alpha^2\sigma^2}{n} \\
&= \|\overline{\mathbf{x}}_k - \mathbf{z}^*\|_2^2 - 2\alpha\langle \overline{\mathbf{x}}_k - \mathbf{z}^*, \overline{\mathbf{p}}_k\rangle + 2\alpha\langle \overline{\mathbf{x}}_k - \mathbf{z}^*, \overline{\mathbf{p}}_k - \overline{\mathbf{h}}_k\rangle + \alpha^2\|\overline{\mathbf{p}}_k - \overline{\mathbf{h}}_k\|_2^2 \\
&\quad + \alpha^2\|\overline{\mathbf{p}}_k\|_2^2 - 2\alpha^2\langle \overline{\mathbf{p}}_k, \overline{\mathbf{p}}_k - \overline{\mathbf{h}}_k\rangle + \frac{\alpha^2\sigma^2}{n} \\
&= \|\overline{\mathbf{x}}_k - \alpha\overline{\mathbf{p}}_k - \mathbf{z}^*\|_2^2 + \alpha^2\|\overline{\mathbf{p}}_k - \overline{\mathbf{h}}_k\|_2^2 + 2\alpha\langle \overline{\mathbf{x}}_k - \alpha\overline{\mathbf{p}}_k - \mathbf{z}^*, \overline{\mathbf{p}}_k - \overline{\mathbf{h}}_k\rangle + \frac{\alpha^2\sigma^2}{n} \\
&\leq \eta^2\|\overline{\mathbf{x}}_k - \mathbf{z}^*\|_2^2 + \alpha^2\|\overline{\mathbf{p}}_k - \overline{\mathbf{h}}_k\|_2^2 + 2\alpha\eta\|\overline{\mathbf{x}}_k - \mathbf{z}^*\|_2\|\overline{\mathbf{p}}_k - \overline{\mathbf{h}}_k\|_2 + \frac{\alpha^2\sigma^2}{n} \\
&\leq (1-\alpha\mu)\|\overline{\mathbf{x}}_k - \mathbf{z}^*\|_2^2 + \left(\frac{\alpha\ell^2}{n\mu}\right)(1+\alpha\mu)\|\mathbf{1}_n\overline{\mathbf{x}}_k - \mathbf{z}_k\|_2^2 + \frac{\alpha^2\sigma^2}{n}.
\end{aligned} \tag{20}$$

It can be verified that $B^\infty = \frac{1}{n} Y^\infty \mathbf{1}_n \mathbf{1}_n^\top$. Next consider $\|\mathbf{z}_k - \mathbf{1}_n \overline{\mathbf{x}}_k\|_2^2$:

$$\begin{aligned}
\|\mathbf{z}_k - \mathbf{1}_n\overline{\mathbf{x}}_k\|_2^2 &= \|Y^{-1}\mathbf{x}_k - Y^\infty \mathbf{1}_n \overline{\mathbf{x}}_k + Y^\infty \mathbf{1}_n \overline{\mathbf{x}}_k - \mathbf{1}_n\overline{\mathbf{x}}_k\|_2^2 \\
&= \|Y^{-1}(\mathbf{x}_k - Y^\infty \mathbf{1}_n\overline{\mathbf{x}}_k) + (Y^{-1}Y^\infty - I_n)\mathbf{1}_n\overline{\mathbf{x}}_k\|_2^2 \\
&= \|Y^{-1}(\mathbf{x}_k - B^\infty \mathbf{x}_k)\|_2^2 + \|(Y^{-1}Y^\infty - I_n)\mathbf{1}_n\overline{\mathbf{x}}_k\|_2^2 + 2\langle Y^{-1}(\mathbf{x}_k - B^\infty\mathbf{x}_k), (Y^{-1}Y^\infty - I_n)\mathbf{1}_n\overline{\mathbf{x}}_k\rangle \\
&\le y_-^2\|\mathbf{x}_k - B^\infty \mathbf{x}_k\|_2^2 + (y_-\beta\sigma_B^k)^2\|\mathbf{x}_k\|_2^2 + 2(y_-)(y_-\beta\sigma_B^k)\|\mathbf{x}_k - B^\infty\mathbf{x}_k\|_2\|\mathbf{x}_k\|_2 \\
&\le (y_-^2 + y_-^2\beta\sigma_B)\overline{\pi}\|\mathbf{x}_k - B^\infty\mathbf{x}_k\|_{\boldsymbol{\pi}}^2 + \left(y_-^2\beta^2\sigma_B^{2k} + y_-^2\beta\sigma_B^k\right)\|\mathbf{x}_k\|_2^2.
\end{aligned}$$

Using the above relation in (20), we obtain the final expression for $\mathbb{E}\left[\|\overline{\mathbf{x}}_{k+1} - \mathbf{z}^*\|_2^2 | \mathcal{F}_k\right]$.

$$\begin{aligned}
\mathbb{E}\left[\|\overline{\mathbf{x}}_{k+1} - \mathbf{z}^*\|_2^2|\mathcal{F}_k\right] &\le (\alpha^2 g_1 + \alpha g_2)\|\mathbf{x}_k - B^\infty\mathbf{x}_k\|_{\boldsymbol{\pi}}^2 + (1-\alpha\mu)\|\overline{\mathbf{x}}_k - \mathbf{z}^*\|_2^2 \\
&+ \alpha^2\left(\frac{\sigma^2}{n}\right) + (h_1\sigma_B^k)\|\mathbf{x}_k\|_2^2. \qquad (21)
\end{aligned}$$

Step 3: Gradient tracking error. Finally, we calculate the gradient tracking error $\|\mathbf{w}_{k+1} - B^\infty \mathbf{w}_{k+1}\|_{\boldsymbol{\pi}}^2$.

$$\begin{aligned}
\|\mathbf{w}_{k+1} - B^\infty\mathbf{w}_{k+1}\|_{\boldsymbol{\pi}}^2 &= \|B\mathbf{w}_k - B^\infty\mathbf{w}_k + (I_n - B^\infty)(\nabla\widehat{f}(\mathbf{z}_{k+1}) - \nabla\widehat{f}(\mathbf{z}_k))\|_{\boldsymbol{\pi}}^2 \\
&\le \sigma_B^2\|\mathbf{w}_k - B^\infty\mathbf{w}_k\|_{\boldsymbol{\pi}}^2 + \|\!| I_n - B^\infty \|\!|_{\boldsymbol{\pi}}^2 \|\nabla\widehat{f}(\mathbf{z}_{k+1}) - \nabla\widehat{f}(\mathbf{z}_k)\|_{\boldsymbol{\pi}}^2 \\
&+ 2\sigma_B\langle \mathbf{w}_k - B^\infty\mathbf{w}_k, (I_n - B^\infty)(\nabla\widehat{f}(\mathbf{z}_{k+1}) - \nabla\widehat{f}(\mathbf{z}_k))\rangle_{\boldsymbol{\pi}} \\
&\le \sigma_B^2\|\mathbf{w}_k - B^\infty\mathbf{w}_k\|_{\boldsymbol{\pi}}^2 + \|\nabla\widehat{f}(\mathbf{z}_{k+1}) - \nabla\widehat{f}(\mathbf{z}_k)\|_{\boldsymbol{\pi}}^2 \\
&+ 2\sigma_B\|\mathbf{w}_k - B^\infty\mathbf{w}_k\|_{\boldsymbol{\pi}} \|\!| I_n - B^\infty \|\!|_{\boldsymbol{\pi}} \|\nabla\widehat{f}(\mathbf{z}_{k+1}) - \nabla\widehat{f}(\mathbf{z}_k))\|_{\boldsymbol{\pi}} \\
&\le \left(\sigma_B^2 + \sigma_B\frac{1-\sigma_B^2}{2\sigma_B}\right)\|\mathbf{w}_k - B^\infty\mathbf{w}_k\|_{\boldsymbol{\pi}}^2 + \left(1 + \sigma_B\frac{2\sigma_B}{1-\sigma_B^2}\right)\|\nabla\widehat{f}(\mathbf{z}_{k+1}) - \nabla\widehat{f}(\mathbf{z}_k)\|_{\boldsymbol{\pi}}^2 \\
&= \left(\frac{1+\sigma_B^2}{2}\right)\|\mathbf{w}_k - B^\infty\mathbf{w}_k\|_{\boldsymbol{\pi}}^2 + \left(\frac{1+\sigma_B^2}{1-\sigma_B^2}\right)\|\nabla\widehat{f}(\mathbf{z}_{k+1}) - \nabla\widehat{f}(\mathbf{z}_k)\|_{\boldsymbol{\pi}}^2.
\end{aligned}$$

We bound the second term of the above equation as:

$$\begin{aligned}
\|\nabla\widehat{f}(\mathbf{z}_{k+1}) - \nabla\widehat{f}(\mathbf{z}_k)\|_{\boldsymbol{\pi}}^2 &= \|\nabla\widehat{f}(\mathbf{z}_{k+1}) - \nabla\widehat{f}(\mathbf{z}_k) - (\nabla f(\mathbf{z}_{k+1}) - \nabla f(\mathbf{z}_k)) + \nabla f(\mathbf{z}_{k+1}) - \nabla f(\mathbf{z}_k)\|_{\boldsymbol{\pi}}^2 \\
&\le 2\ell^2\underline{\pi}^{-1}\|\mathbf{z}_{k+1} - \mathbf{z}_k\|_2^2 + 2\|\nabla\widehat{f}(\mathbf{z}_{k+1}) - \nabla\widehat{f}(\mathbf{z}_k) - (\nabla f(\mathbf{z}_{k+1}) - \nabla f(\mathbf{z}_k))\|_{\boldsymbol{\pi}}^2.
\end{aligned}$$

Consider the first term $\|\mathbf{z}_{k+1} - \mathbf{z}_k\|_2^2$ of above equation.

$$\begin{aligned}
\|\mathbf{z}_{k+1} - \mathbf{z}_k\|_2^2 &= \|Y_{k+1}^{-1}((B\mathbf{x}_k - \alpha\mathbf{w}_k) - \mathbf{x}_k) + (Y_{k+1}^{-1} - Y_k^{-1})\mathbf{x}_k\|_2^2 \\
&= \|Y_{k+1}^{-1}(B - I_n)\mathbf{x}_k - \alpha Y_{k+1}^{-1}\mathbf{w}_k + (Y_{k+1}^{-1} - Y_k^{-1})\mathbf{x}_k\|_2^2 \\
&\le \|Y_{k+1}^{-1}(B - I_n)\mathbf{x}_k\|_2^2 + \alpha^2\|Y_{k+1}^{-1}\mathbf{w}_k\|_2^2 + \|(Y_{k+1}^{-1} - Y_k^{-1})\mathbf{x}_k\|_2^2 + 2\|Y_{k+1}^{-1}(B - I_n)\mathbf{x}_k\|_2\|\alpha Y_{k+1}^{-1}\mathbf{w}_k\|_2 \\
&+ 2\|\alpha Y_{k+1}^{-1}\mathbf{w}_k\|_2\|(Y_{k+1}^{-1} - Y_k^{-1})\mathbf{x}_k\|_2 + 2\|Y_{k+1}^{-1}(B - I_n)\mathbf{x}_k\|_2\|(Y_{k+1}^{-1} - Y_k^{-1})\mathbf{x}_k\|_2 \\
&\le \|Y_{k+1}^{-1}(B - I_n)\mathbf{x}_k\|_2^2 + \|\alpha Y_{k+1}^{-1}\mathbf{w}_k\|_2^2 + \left\|\!\left| Y_{k+1}^{-1} - Y_k^{-1} \right|\!\right\|_2^2\|\mathbf{x}_k\|_2^2 + 2\|Y_{k+1}^{-1}(B - I_n)\mathbf{x}_k\|_2\|\alpha Y_{k+1}^{-1}\mathbf{w}_k\|_2 \\
&+ 2\alpha\|Y_{k+1}^{-1}\mathbf{w}_k\|_2\left\|\!\left| Y_{k+1}^{-1} - Y_k^{-1} \right|\!\right\|_2\|\mathbf{x}_k\|_2 + 2\|Y_{k+1}^{-1}(B - I_n)\mathbf{x}_k\|_2\left\|\!\left| Y_{k+1}^{-1} - Y_k^{-1} \right|\!\right\|_2\|\mathbf{x}_k\|_2 \\
&\le 12y_-^2\overline{\pi}\|\mathbf{x}_k - B^\infty\mathbf{x}_k\|_{\boldsymbol{\pi}}^2 + 3\alpha^2y_-^2\|\mathbf{w}_k\|_2^2 + 24y_-^4\beta^2\sigma_B^{2k}\|\mathbf{x}_k\|_2^2.
\end{aligned}$$

Next we bound $\|\mathbf{w}_k\|_2^2$,

$$\begin{aligned}\|\mathbf{w}_k\|_2^2 &= \|(\mathbf{w}_k - Y^\infty \mathbf{1}_n \overline{\mathbf{g}}_k) + Y^{-1}Y^\infty \mathbf{1}_n \overline{\mathbf{p}}_k + Y^{-1}Y^\infty(\mathbf{1}_n\overline{\mathbf{g}}_k - \mathbf{1}_n\overline{\mathbf{p}}_k)\|_2^2 \\ &\leq (2+r)\|\mathbf{w}_k - Y^\infty \mathbf{1}_n \overline{\mathbf{w}}_k\|_2^2 + 3\|Y^{-1}Y^\infty \mathbf{1}_n \overline{\mathbf{p}}_k\|_2^2 + \left(2+\frac{1}{r}\right)\|Y^{-1}Y^\infty \mathbf{1}_n(\overline{\mathbf{g}}_k - \overline{\mathbf{p}}_k)\|_2^2 \\ &\leq (2+r)\overline{\pi}\|\mathbf{w}_k - B^\infty \mathbf{w}_k\|_{\boldsymbol{\pi}}^2 + 3y_-^2 y^2 \ell^2 \|\overline{\mathbf{x}}_k - \mathbf{z}^*\|_2^2 + 2\left(2+\frac{1}{r}\right) y_-^2 y^2 n \|\overline{\mathbf{g}}_k - \overline{\mathbf{h}}_k\|_2^2 \\ &+ 2\left(2+\frac{1}{r}\right) y_-^2 y^2 \ell^2 \|\mathbf{z}_k - \mathbf{1}_n \overline{\mathbf{x}}_k\|_2^2.\end{aligned}$$

whereas,

$$\mathbb{E}[\|\nabla \widehat{f}(\mathbf{z}_{k+1}) - \nabla \widehat{f}(\mathbf{z}_k) - (\nabla f(\mathbf{z}_{k+1}) - \nabla f(\mathbf{z}_k))\|_{\boldsymbol{\pi}}^2 | \mathcal{F}_k] = 2n\sigma^2 \underline{\pi}^{-1}.$$

Pick $r = \frac{k_1}{k_2\alpha^2} - 2 = \frac{k_1 - 2k_2\alpha^2}{k_2\alpha^2} > 0 \Rightarrow \frac{1}{r} = \frac{k_2\alpha^2}{k_1 - 2k_2\alpha^2} > 0$. This will enforce a constraint on α such that $\alpha < \sqrt{\frac{k_1}{2k_2}} = \left(\frac{1-\sigma_B^2}{6\ell y_-}\right)\sqrt{\frac{\underline{\pi}}{(1+\sigma_B^2)\overline{\pi}}}$. The term $\|\mathbf{z}_k - \mathbf{1}_n \overline{\mathbf{x}}_k\|_2^2$ is already simplified in solving for the optimality gap. Putting these in above equation and after taking the expectation, the resultant equation for gradient tracking error becomes:

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{w}_{k+1} - B^\infty \mathbf{w}_{k+1}\|_{\boldsymbol{\pi}}^2 | \mathcal{F}_k\right] &\leq (g_3 + \alpha^2 g_4)\|\mathbf{x}_k - B^\infty \mathbf{x}_k\|_{\boldsymbol{\pi}}^2 + (\alpha^2 g_5)\|\overline{\mathbf{x}}_k - \mathbf{z}^*\|_2^2 + C_\sigma \\ &+ \left(\frac{5+\sigma_B^2}{6}\right)\mathbb{E}\left[\|\mathbf{w}_k - B^\infty \mathbf{w}_k\|_{\boldsymbol{\pi}}^2 | \mathcal{F}_k\right] + ((h_2 + \alpha^2 h_3)\sigma_B^k)\|\mathbf{x}_k\|_2^2. \end{aligned} \tag{22}$$

Taking full expectation of (19), (21), and (22) leads to the system dynamics described by the relation in (7).

Appendix B

Proof of Corollary 1

We derive the upper bound on the spectral radius of A_α under the conditions on step-size described in Theorem 1. Using (8) and (9), the characteristic function of A_α can be calculated as:

$$\begin{aligned}\det(\lambda I_3 - A_\alpha) &= (\lambda - a_{11})(\lambda - a_{22})(\lambda - a_{33}) - a_{13}a_{31}(\lambda - a_{22}) - a_{13}a_{21}a_{32} \\ &\geq (\lambda - a_{11})(\lambda - a_{22})(\lambda - a_{33}) - a_{13}a_{31}(\lambda - a_{22}) - \frac{1}{\Gamma+1}(1-a_{22})[(1-a_{11})(1-a_{33}) - a_{13}a_{31}] \\ &\geq (\lambda - a_{11})(\lambda - a_{22})(\lambda - a_{33}) - \frac{1}{\Gamma}(\lambda - a_{22})(1-a_{11})(1-a_{33}) \\ &- \frac{\Gamma - 1}{\Gamma(\Gamma+1)}(1-a_{11})(1-a_{22})(1-a_{33}).\end{aligned}$$

Since the $\det(\lambda I - A_\alpha) > 0$ and the $\det(\max\{a_{11}, a_{22}, a_{33}\}I - A_\alpha) = \det(a_{22}I - A_\alpha) < 0$, the spectral radius $\rho(A_\alpha) = (a_{22}, 1)$. Suppose $\lambda = 1 - \epsilon$ for some $\epsilon \in (0, \alpha\mu)$, satisfying

$$\begin{aligned}
\det(\lambda I_3 - A_\alpha) \geq & \left(1 - \epsilon - \frac{1+\sigma_B^2}{2}\right)(\alpha\mu - \epsilon)\left(1 - \epsilon - \frac{5+\sigma_B^2}{6}\right) - \frac{1}{\Gamma}(\alpha\mu - \epsilon)\left(1 - \frac{1+\sigma_B^2}{2}\right)\left(1 - \frac{5+\sigma_B^2}{6}\right) \\
& - \frac{\Gamma - 1}{\Gamma(\Gamma+1)}\left(1 - \frac{1+\sigma_B^2}{2}\right)(\alpha\mu)\left(1 - \frac{5+\sigma_B^2}{6}\right) \geq 0, \\
\iff & \left(\frac{1-\sigma_B^2 - 2\epsilon}{2}\right)(\alpha\mu - \epsilon)\left(\frac{1-\sigma_B^2 - 6\epsilon}{6}\right) - \frac{1}{\Gamma}(\alpha\mu - \epsilon)\left(\frac{1-\sigma_B^2}{2}\right)\left(\frac{1-\sigma_B^2}{6}\right) \\
& - \frac{\Gamma - 1}{\Gamma(\Gamma+1)}\left(\frac{1-\sigma_B^2}{2}\right)(\alpha\mu)\left(\frac{1-\sigma_B^2}{6}\right) \geq 0, \\
\iff & (\alpha\mu - \epsilon)\left[(1-\sigma_B^2 - 2\epsilon)(1-\sigma_B^2 - 6\epsilon) - \frac{1}{\Gamma}(1-\sigma_B^2)^2\right] \geq \frac{\Gamma - 1}{\Gamma(\Gamma+1)}(1-\sigma_B^2)^2(\alpha\mu), \\
\iff & \left(\frac{\alpha\mu - \epsilon}{\alpha\mu}\right)\left[\frac{(1-\sigma_B^2 - 2\epsilon)(1-\sigma_B^2 - 6\epsilon)}{(1-\sigma_B^2)^2} - \frac{1}{\Gamma}\right] \geq \frac{\Gamma - 1}{\Gamma(\Gamma+1)}. \qquad (23)
\end{aligned}$$

It is sufficient to have

$$\epsilon \leq \left(\frac{\Gamma - 1}{\Gamma + 1}\right)\alpha\mu.$$

Notice that,

$$\left(\frac{\alpha\mu - \epsilon}{\alpha\mu}\right) \geq \left(\frac{\alpha\mu - \left(\frac{\Gamma-1}{\Gamma+1}\right)\alpha\mu}{\alpha\mu}\right) = 1 - \left(\frac{\Gamma - 1}{\Gamma + 1}\right) = \frac{\Gamma + 1 - \Gamma + 1}{\Gamma + 1} = \frac{2}{\Gamma + 1}.$$

To verify the upper bound on ϵ under the condition on step-size described in Corollary 1,

$$\epsilon \leq \left(\frac{\Gamma - 1}{\Gamma + 1}\right)\left(\frac{\Gamma + 1}{\Gamma}\right)\left(\frac{1-\sigma_B^2}{20\mu}\right)\mu = \left(\frac{\Gamma - 1}{\Gamma}\right)\left(\frac{1-\sigma_B^2}{20}\right),$$

which implies,

$$\begin{aligned}
1 - \sigma_B^2 - 2\epsilon \geq 1 - \sigma_B^2 - 2\left(\frac{\Gamma - 1}{\Gamma}\right)\left(\frac{1-\sigma_B^2}{20}\right) &= \frac{(9\Gamma + 1)(1-\sigma_B^2)}{10\Gamma}, \\
1 - \sigma_B^2 - 6\epsilon \geq 1 - \sigma_B^2 - 6\left(\frac{\Gamma - 1}{\Gamma}\right)\left(\frac{1-\sigma_B^2}{20}\right) &= \frac{(7\Gamma + 3)(1-\sigma_B^2)}{10\Gamma}, \\
\iff (1-\sigma_B^2 - 2\epsilon)(1-\sigma_B^2 - 6\epsilon) &\geq \frac{(63\Gamma^2 + 34\Gamma + 3)(1-\sigma_B^2)^2}{100\Gamma^2}.
\end{aligned}$$

Plugging these values in (23) and for $\Gamma > 1$, we get,

$$\begin{aligned}
\left(\frac{\alpha\mu - \epsilon}{\alpha\mu}\right)\left[\frac{(1-\sigma_B^2 - 2\epsilon)(1-\sigma_B^2 - 6\epsilon)}{(1-\sigma_B^2)^2} - \frac{1}{\Gamma}\right] &\geq \left(\frac{2}{\Gamma + 1}\right)\left[\frac{\frac{(63\Gamma^2 + 34\Gamma + 3)(1-\sigma_B^2)^2}{100\Gamma^2}}{(1-\sigma_B^2)^2} - \frac{1}{\Gamma}\right] \\
&= \left(\frac{1}{\Gamma(\Gamma+1)}\right)\left[\frac{(63\Gamma^2 + 34\Gamma + 3)}{50\Gamma} - 2\right] = \left(\frac{1}{\Gamma(\Gamma+1)}\right)\left[\frac{63\Gamma^2 - 66\Gamma + 3}{50\Gamma}\right] \\
&= \left(\frac{1}{\Gamma(\Gamma+1)}\right)\left[\Gamma - 1 + \frac{13\Gamma}{50} - \frac{16}{50} + \frac{3}{50\Gamma}\right] \geq \frac{\Gamma - 1}{\Gamma(\Gamma+1)}.
\end{aligned}$$

Define $\lambda^* = 1 - \left(\frac{\Gamma-1}{\Gamma+1}\right)\alpha\mu$. Then the $\det(\lambda^* I - A_\alpha) \geq 0$. Therefore, $\rho(A_\alpha) \leq \lambda^*$. We select $\Gamma = 2$ and the corollary follows.