Accelerated Distributed Stochastic Non-Convex Optimization over Time-Varying Directed Networks

Yiyue Chen, *Student Member, IEEE*, Abolfazl Hashemi, *Member, IEEE*, and Haris Vikalo, *Senior Member, IEEE*

***Abstract*—Distributed stochastic non-convex optimization problems have recently received attention due to the growing interest of signal processing, computer vision, and natural language processing communities in applications deployed over distributed learning systems (e.g., federated learning). We study the setting where the data is distributed across the nodes of a time-varying directed network, a topology suitable for modeling dynamic networks experiencing communication delays and straggler effects. The network nodes, which can access only their local objectives and query a stochastic first-order oracle to obtain gradient estimates, collaborate to minimize a global objective function by exchanging messages with their neighbors. We propose an algorithm, novel to this setting, that leverages stochastic gradient descent with momentum and gradient tracking to solve distributed non-convex optimization problems over time-varying networks. To analyze the algorithm, we tackle the challenges that arise when analyzing dynamic network systems which communicate gradient acceleration components. We prove that the algorithm's oracle complexity is $\mathcal{O}(1/\epsilon^{1.5})$, and that under Polyak-Łojasiewicz condition the algorithm converges linearly to a steady error state. The proposed scheme is tested on several learning tasks: a non-convex logistic regression experiment on the MNIST dataset, an image classification task on the CIFAR-10 dataset, and an NLP classification test on the IMDB dataset. We further present numerical simulations with an objective that satisfies the PL condition. The results demonstrate superior performance of the proposed framework compared to the existing related methods.**

***Index Terms*—decentralized non-convex optimization, stochastic non-convex optimization, time-varying directed network**

I. Introduction

We study distributed non-convex optimization problems encountered in a variety of applications in machine learning (ML), signal processing, and control [2, 3, 4]. Distributed learning frameworks aim to address limitations of centralized methods including the potentially high cost of communicating data to a central location, privacy and latency concerns, and data storage constraints [5]. We model a distributed computing system via a time-varying directed network $\mathcal{G}(t) = (\mathcal{V}, \mathcal{E}(t))$, where $\mathcal{V} = \{1, \cdots, n\}$ denotes the set of n nodes and $\mathcal{E}(t)$ is the collection of directed edges (i, j), $i, j \in \mathcal{V}$, connecting the nodes at time t. In particular, if $(i, j) \in \mathcal{E}(t)$, there exists an edge from node i to node j, and thus node i can send messages to node j at time t. Node i has access only to its local data and the local loss function. The goal of the nodes in the network is to collaboratively minimize a global loss function, i.e., solve

$$\min_{\mathbf{x}\in\mathbb{R}^d} \left[f(\mathbf{x}) := \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}) \right], \tag{1}$$

where $f_i(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$ for $i \in [n] := \{1, ..., n\}$ denotes the non-convex objective that the device at node i minimizes locally; in the machine learning context, this describes the setting where each node trains a local model by optimizing a cost function with d parameters and collaborates with other nodes to find the global model.

In a departure from the existing work focused on *undirected networks* [6, 7, 8, 9], we study distributed non-convex optimization over *time-varying directed networks* where each node minimizes its local objective utilizing a stochastic gradient obtained by querying a local stochastic first-order oracle, i.e., the nodes use noisy estimates of the local gradient at the query point. Unlike undirected networks, directed communication topologies take into account a number of practical considerations including asymmetry in the communication links (e.g., in the multi-agent applications) and the straggler effects stemming from imposing synchronized communication. Furthermore, time-varying networks characterize the communication link delay or failure in real-world applications.

A. Related work and significance

Decentralized non-convex optimization problems with stochastic first-order oracles have been extensively studied in the context of undirected networks, where doubly stochastic weight matrices lead to convergence guarantees. Those studies include the decentralized stochastic gradient descent (DSGD) and its variants, [14, 15, 6], which combine decentralized average consensus with local gradient updates. Although DSGD is often effective and relatively simple to implement, it is unstable in settings that involve heterogeneous data [7]. This has motivated the search for more robust schemes which combine decentralized bias-correction techniques, gradient tracking, and primal-dual methods [16, 10, 9, 17]. Building on top of such techniques, GT-HSGD [8] leverages SARAH-type variance reduction schemes (see, e.g., [18, 19]) to further reduce oracle complexities under the so-called mean-squared smoothness [20]. However, GT-HSGD still relies on doubly stochastic weight matrices hindering its practical feasibility in applications involving asymmetric communication.

While there has been extensive prior work on distributed optimization over the family of networks characterized by doubly-stochastic mixing matrix, e.g., undirected networks and some special cases of directed networks, practical systems experience transmission failures and/or bandwidth limitations which imply asymmetric or uni-directional communication between network nodes. In such scenarios, more general directed graphs that do not satisfy doubly-stochastic property are a better-suited network model [5]; however, the design of convergent algorithms for distributed optimization over general directed graphs brings forth new challenges. In particular, we recall that to ensure convergence of the algorithms for decentralized optimization over undirected networks, the weight (mixing) matrix W_m should be symmetric and doubly stochastic. Indeed, mixing matrix characterizes communication over a network: when doubly stochastic, the decentralized algorithm reaches the average consensus model since $\lim_{T\to\infty} \Pi_{t=1}^{T} W_m = \frac{\mathbf{1}\mathbf{1}^T}{n}$. When a network is directed, however, the communication links are asymmetric and the corresponding mixing matrix is generally not doubly stochastic. There, it holds that $\lim_{T\to\infty} \Pi_{t=1}^{T} W_m = \pi\mathbf{1}^T$ but unlike the undirected network scenario $\pi_i \neq \frac{1}{n}$, which implies convergence to a biased weighted average of local models. As a remedy, distributed optimization schemes for directed networks often deploy auxiliary variables to help deal

This work was supported in part by NSF grants 1809327 and CNS-2313109.

Yiyue Chen and Haris Vikalo are with the Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX 78712 USA. Abolfazl Hashemi is with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA.

A preliminary version of this article is presented at the 2023 International Conference on Acoustics, Speech, and Signal Processing (ICASSP) [1].

TABLE I

A comparison of algorithms for decentralized optimization over directed graphs. SFO and IFO stand for stochastic and incremental first-order oracles, respectively. SGP relies on a bounded dissimilarity assumption to obtain the stated result. GT-HSGD utilizes the non-standard doubly-stochastic requirement for the weight matrix. The stated results for SGP and GT-HSGD are semi-asymptotic, meaning they require a large number of iterations to achieve the stated convergence bounds. Double-stochasticity is satisfied by all undirected networks and a small collection of directed networks; 'dynamic network' refers to time-varying networks, i.e., indicates that the network is changing over time.

Algorithm	Double-stochasticity	Dynamic network	Oracle Complexity	Remarks
SGP/Push-SGD [5]	No	Yes	$\mathcal{O}(\frac{1}{n\epsilon^2})$	SFO, bounded dissimilarity only for large T
Push-DIGing [10]	No	Yes	$\mathcal{O}(\ln\frac{1}{\epsilon})$	deterministic Strong convexity, smoothness
Di-CS-SVRG [11]	No	Yes	$\mathcal{O}(\ln\frac{1}{\epsilon})$	IFO, strong convexity smoothness
Push-SAGA [12]	No	No	$\mathcal{O}(\ln\frac{1}{\epsilon})$	IFO, strong convexity smoothness
S-ADDOPT [13]	No	No	$\mathcal{O}(\frac{1}{\epsilon})$	SFO, strong convexity smoothness
GT-HSGD [8]	Yes	No	$\mathcal{O}(\frac{1}{n\epsilon^{1.5}})$	SFO, mean-squared smoothness doubly stochastic weight matrix only for large T
Push-ASGD (This work)	No	Yes	$\mathcal{O}(\frac{1}{\epsilon^{1.5}})$	SFO, mean-squared smoothness
Push-ASGD (This work)	No	Yes	$\mathcal{O}(\ln\frac{1}{\epsilon})$	SFO, mean-squared smoothness, PL condition

with the communication asymmetry. For instance, frequently used subgradient-push algorithm [2, 21] and its variants [13], which operates on column-stochastic mixing matrices, introduces local normalization scalars to de-bias the weighted average and thus ensure convergence. In another line of related work, [22, 23] introduce auxiliary variables of the same dimension as the local model parameters to keep track of the local parameter variations and avoid division (a nonlinear operation) deployed by the subgradient-push algorithm. When the objective is smooth and strongly convex, a linear rate can be achieved by using constant step size [10, 24, 11, 25]. Further acceleration of the convergence rate can be achieved adopting momentum-based methods [26].

Distributed, stochastic *non-convex* optimization over directed time-varying graphs has received relatively little attention [5]. Even though some of the above optimization algorithms (developed with convex objectives in mind) can be applied to distributed non-convex optimization problems, they either converge at a slow rate or apply without any theoretical guarantees. Aiming to achieve robust and provably fast performance, we present and analyze an algorithm that relies on gradient-push, global gradient tracking, and a local hybrid gradient estimator with momentum to solve distributed, stochastic non-convex optimization problems that arise in modern distributed ML tasks. Our contributions are summarized as follows:

1. We study (to our knowledge, previously not pursued in literature) the problem of distributed non-convex optimization under a stochastic first-order oracle (SFO) over directed time-varying networks. We propose for this setting novel variance-reduced algorithm, Push-ASGD, which leverages gradient-push, global gradient tracking, and hybrid gradient estimation methods. In particular, we devise a new stochastic gradient estimator used locally by each node, and design a framework that operates with column stochastic weight matrices.

2. To analyze convergence of Push-ASGD, we address challenges brought forward by the exchange of gradient acceleration components across directed time-varying networks. We prove that, under the mean-squared smoothness, the algorithm attains an ϵ-accurate first-order stationary solution with an oracle complexity of $\mathcal{O}(1/\epsilon^{1.5})$. To our knowledge, this is the first algorithm to come with such guarantees in the context of stochastic nonconvex optimization over directed networks. We further show that for objective functions satisfying Polyak-Łojasiewicz (PL) condition, when using constant step size the algorithm converges linearly to a steady state with small error.

3. We validate the proposed algorithm on various distributed learning problems including image classification and natural language processing via deep learning, demonstrating its superior accuracy/convergence compared to relevant existing techniques. We also test its performance in simulations involving an objective function that satisfies the PL condition. The results demonstrate superior accuracy/convergence of Push-ASGD compared to the relevant benchmarking algorithms.

II. Preliminaries

Assume that n nodes are collaboratively solving the decentralized non-convex optimization problem (1) while communicating over a time-varying directed network. Each node i in the network trains a local model and computes a sequence of local model estimates $\{\mathbf{x}_t^i\}$ towards a first-order stationary point of the global objective f, starting from a pre-specified initialization point $\mathbf{x}_0^i$. To update its local model estimate, node i accesses a random local data point ξ_t^i and queries stochastic first-order oracle for a stochastic gradient, $\nabla f_i(\mathbf{x}, \xi_t^i)$, given the input $\mathbf{x}$. To characterize the sequence of random local data points, we consider the filtration induced by the data points of all the nodes in the network,

$$\mathcal{F}_0 = \{\Omega, \phi\}, \quad \mathcal{F}_t = \sigma(\{\xi_0^i, \xi_1^i, \cdots, \xi_{t-1}^i, i \in \mathcal{V}\}), \forall t \geq 1, \quad (2)$$

where ϕ is the empty set and $\{\mathcal{F}_t\}$ is an increasing family of σ-algebras. The input vector $\mathbf{x}$ at iteration t is $\mathcal{F}_t$-measurable. We denote the probability space by $\{\Omega, \mathbb{P}, \mathcal{F}\}$. We make the following assumptions on the stochastic first-order oracles, the global objective, and the communication network.

Assumption 1. *For all $i \in \mathcal{V}$ and all $t \geq 0$, we assume:*
1. Unbiasedness of the conditional expectation

$$\mathbb{E}[\nabla f_i(\mathbf{x}, \xi_t^i)|\mathcal{F}_t] = \nabla f_i(\mathbf{x}).$$

2. Bounded variance of the estimated gradient,

$$\mathbb{E}\|\nabla f_i(\mathbf{x}, \xi_t^i) - \nabla f_i(x)\|^2 \leq \nu_i^2.$$

It will be convenient to introduce $\bar{\nu}^2 = \frac{1}{n}\sum_{i=1}^n \nu_i^2$.
3. Independent random selection, i.e., the random vectors $\{\xi_0^i, \xi_1^i, \cdots, \xi_{t-1}^i\}$, $i \in \mathcal{V}$, are independent.
4. The mean-squared smoothness,

$$\mathbb{E}\|\nabla f_i(\mathbf{x}, \xi_t^i) - \nabla f_i(\mathbf{y}, \xi_t^i)\|^2 \leq L^2 \mathbb{E}\|\mathbf{x} - \mathbf{y}\|^2.$$

The first three assumptions are widely used in the analysis of stochastic first-order optimization algorithms [27]. The last assumption requires L-smoothness of the stochastic gradient on average with respect to any two inputs. The mean-squared smoothness further implies smoothness of each local objective f_i, and consequently implies L-smoothness of the global objective [20, 28].

Assumption 2. *The global objective is lower bounded,*

$$f^* = \inf_{\mathbf{x}} f(\mathbf{x}) > -\infty.$$

Assumption 3. *The network is directed and time-varying. At time t, the network is strongly-connected with a column stochastic weight matrix $W_m^{(t)}$.*

The above assumption on network topology is standard in distributed optimization [10] and more general than the static network assumption in [5]. Elaborating on Assumption 3, let us consider the mixing matrix $W_m^{(t)} := [w_{ij}^{(t)}]$ which captures properties of the communication links in the network; $w_{ij}^{(t)} = \frac{1}{d_j^{out,t}+1} > 0$ if and only if $(j, i) \in \mathcal{E}(t)$ or $i = j$, where $d_j^{out,t}$ is the out-degree of agent j at time t. We assume that node i knows which nodes it sends messages to, i.e., the i-th column of $W_m^{(t)}$ are known to node i. The column stochastic mixing matrix $W_m^{(t)}$ has the left eigenvector $\mathbf{1}_n$ and a positive right eigenvector $\pi^{(t)}$, i.e., $\mathbf{1}_n^\top W_m^{(t)} = \mathbf{1}_n^\top$ and $W_m^{(t)}\pi^{(t)} = \pi^{(t)}$.

III. The Push-ASGD Algorithm

In this section we present an algorithm for distributed non-convex optimization over directed time-varying networks. At a high level, the algorithm relies on the push-sum protocol [2] to perform average consensus, and deploys a stochastic gradient estimator of the unknown global gradient while simultaneously reducing the variance/noise in the local updates via momentum.

At the beginning, all the nodes use the same initial model $\bar{\mathbf{x}}_0$. At iteration t, node i updates its local model $\mathbf{x}_t^i$ by fusing the messages $\mathbf{x}_t^j$ received from its neighbors according to

$$\mathbf{x}_{t+1}^i = \sum_{j=1}^n w_{ij}^{(t)}(\mathbf{x}_t^j - \alpha \mathbf{g}_t^j), \tag{3}$$

where $\mathbf{g}_t^j$ denotes the local stochastic gradient estimate specified below. Since $W_m^{(t)}$ is column-stochastic, the product of $W_m^{(t)}$ over a time duration s, $\Pi_{k=1}^s W_m^{(t+k)}$, generally differ from $\frac{\mathbf{1}\mathbf{1}^T}{n}$, biasing each node to a different model. Therefore, following [29], for each node i with local model $\mathbf{x}_t^i$ at time t we introduce an auxiliary scalar y_t^i and compute a recovering model $\mathbf{z}_t^i = \mathbf{x}_t^i / y_t^i$, enabling de-biasing the local model fused by the mixing matrix. In other words, while the average of $\mathbf{x}_t^i$ is not preserved due to directed communication, the average of the de-biased quantities $\mathbf{z}_t^i$ will be preserved [21].

Algorithm 1 Push Accelerated Stochastic Gradient Descent Algorithm (Push-ASGD)

```
Input: Initialize $\mathbf{x}_0^i = \bar{\mathbf{x}}_0$; $y_0^i = 1$ ; $\mathbf{z}_0^i = \mathbf{x}_0^i$; step size $\alpha$; $\beta \in (0,1)$; time-varying column-stochastic mixing matrix $W_m^{(t)} := [w_{ij}^{(t)}]$; $T \in \mathbf{Z}^+$
Sample $b$ local data points $\{\xi_{0,r}^i\}_{r=1}^b$ and initialize the gradient $\mathbf{g}_0^i = \frac{1}{b}\sum_{r=1}^b \nabla f_i(\mathbf{z}_0^i, \xi_{0,r}^i)$ and the gradient estimator $\mathbf{v}_0^i = \mathbf{g}_0^i$;
for $t = 0$ to $T - 1$ do
    Update the local estimate of the solution
        $\mathbf{x}_{t+1}^i = \sum_{j=1}^n w_{ij}^{(t)}(\mathbf{x}_t^j - \alpha \mathbf{g}_t^j)$
    Update auxiliary variables
        $y_{t+1}^i = \sum_{j=1}^n w_{ij}^{(t)} y_t^j$, $\mathbf{z}_{t+1}^i = \frac{\mathbf{x}_{t+1}^i}{y_{t+1}^i}$
    Sample $\xi_t^i$ and update the local stochastic gradient estimator
        $\mathbf{v}_{t+1}^i = \nabla f_i(\mathbf{z}_{t+1}^i, \xi_t^i) + (1-\beta)(\mathbf{v}_t^i - \nabla f_i(\mathbf{z}_t^i, \xi_t^i))$
    Update the gradient tracker
        $\mathbf{g}_{t+1}^i = \sum_{j=1}^n w_{ij}^{(t)}(\mathbf{g}_t^j + \mathbf{v}_{t+1}^j - \mathbf{v}_t^j)$
end for
```

In addition to addressing challenges that arise from directed communication, the proposed algorithm deals with two sources of variance that hamper the convergence: a local variance that stems from noise in the local stochastic gradients, and a global variance that stems from the heterogeneity of the nodes' data. To address the former, we rely on momentum-based variance reduction, while to tackle the latter we deploy gradient tracking.

Let $\nabla f_i(\mathbf{z}_{t+1}^i, \xi_t^i)$ and $\nabla f_i(\mathbf{z}_t^i, \xi_t^i)$ denote stochastic gradients obtained after querying the local stochastic first-order oracle with $\mathbf{z}_{t+1}^i$ and $\mathbf{z}_t^i$, respectively. The momentum-type update of gradient $\mathbf{v}_{t+1}^i$ is then found as

$$\mathbf{v}_{t+1}^i = \nabla f_i(\mathbf{z}_{t+1}^i, \xi_t^i) + (1-\beta)(\mathbf{v}_t^i - \nabla f_i(\mathbf{z}_t^i, \xi_t^i)), \tag{4}$$

where β denotes the momentum step size controlling the direction of the gradient adjustment term $\mathbf{v}_t^i - \nabla f_i(\mathbf{z}_t^i, \xi_t^i)$. When $\beta = 1$, (4) reduces to the vanilla stochastic gradient descent, while when $\beta = 0$, (4) reduces to a SARAH-type gradient update [30]; neither can achieve the same oracle complexity as the recursive estimator deploying $\beta \in (0, 1)$. This recursive estimator can reduce the variance of the stochastic gradient estimates in both centralized and distributed optimization problems [18, 31]. Finally, the estimate of the global gradient is found via gradient tracking as

$$\mathbf{g}_{t+1}^i = \sum_{j=1}^n w_{ij}^{(t)}(\mathbf{g}_t^j + \mathbf{v}_{t+1}^j - \mathbf{v}_t^j), \tag{5}$$

where the gradient information from neighboring nodes is used to ensure convergence to the first-order stationary point of the global objective.

The above procedure is formalized as Algorithm 1 and in the remainder of the paper referred to as the Push-ASGD (<u>Push</u> <u>A</u>ccelerated <u>S</u>tochastic <u>G</u>radient <u>D</u>escent) algorithm.

IV. Convergence Analysis

We proceed by analytically showing that Push-ASGD achieves $\mathcal{O}(1/\epsilon^{1.5})$ SFO complexity, and that under the PL condition it linearly converges to a small steady-state error. Below, the first theorem establishes the $\mathcal{O}(1/\epsilon^{1.5})$ complexity while the second theorem establishes linear convergence of Push-ASGD.

For convenience, we re-write the key terms in Algorithm 1 as

$$\begin{aligned}
y_{t+1} &= W_m^{(t)} y_t \\
\mathbf{z}_{t+1,n\times d} &= \tilde{W}_m^{(t)}(\mathbf{z}_{t,n\times d} - \alpha \mathbf{h}_{t,n\times d}) \\
\mathbf{h}_{t+1,n\times d} &= \tilde{W}_m^{(t)} \mathbf{h}_{t,n\times d} + \tilde{W}_m^{(t)} Y_t^{-1}(\mathbf{v}_{t+1,n\times d} - \mathbf{v}_{t,n\times d}),
\end{aligned}$$

where $\tilde{W}_m^{(t)} = Y_{t+1}^{-1} W_m^{(t)} Y_t$, $Y_t = diag(y_t)$ and $\mathbf{h}_{t,n\times d} = Y_t^{-1}\mathbf{g}_{t,n\times d}$. Moreover, $\mathbf{z}_{t,n\times d} = [(\mathbf{z}_t^1)^\top; \cdots; (\mathbf{z}_t^n)^\top]$ (similar for $\mathbf{g}_{t,n\times d}, \mathbf{v}_{t,n\times d}$); note that $n \times d$ in the subscript indicates dimension of a matrix. Finally, $\tilde{W}_m^{(t)}$ is a row-stochastic mixing matrix; there exists a stochastic vector sequence $\{\phi_t\}$ such that $\phi_{t+1}^T \tilde{W}_m^{(t)} = \phi_t^T$.[1]

Before stating the theorem, it will be convenient to introduce the global vectors in $\mathbb{R}^{nd}$

$$\begin{aligned}
&\mathbf{x}_t = [(\mathbf{x}_t^1)^\top \cdots, (\mathbf{x}_t^n)^\top]^\top, \ \ \mathbf{z}_t = [(\mathbf{z}_t^1)^\top \cdots, (\mathbf{z}_t^n)^\top]^\top, \\
&\mathbf{g}_t = [(\mathbf{g}_t^1)^\top \cdots, (\mathbf{g}_t^n)^\top]^\top, \quad \mathbf{v}_t = [(\mathbf{v}_t^1)^\top \cdots, (\mathbf{v}_t^n)^\top]^\top, \\
&\nabla \mathbf{f}(\mathbf{z}_t) = [\nabla f_1(\mathbf{z}_t^1)^\top \cdots \nabla f_n(\mathbf{z}_t^n)]^\top, \ \ \mathbf{h}_t = [(\mathbf{h}_t^1)^\top \cdots (\mathbf{h}_t^n)^\top]^\top,
\end{aligned} \tag{6}$$

as well as the averaged global vectors in $\mathbb{R}^{nd}$,

$$\begin{aligned}
\bar{\mathbf{x}}_t &= \frac{1}{n}[(\sum_j \mathbf{x}_t^j)^\top \cdots (\sum_j \mathbf{x}_t^j)^\top]^\top, \\
\bar{\mathbf{v}}_t &= \frac{1}{n}[(\sum_j \mathbf{v}_t^j)^\top \cdots (\sum_j \mathbf{v}_t^j)^\top]^\top, \\
\hat{\mathbf{z}}_t &= [(\sum_j [\phi_t]_j \mathbf{z}_t^j)^\top \cdots (\sum_j [\phi_t]_j \mathbf{z}_t^j)^\top], \\
\nabla \bar{\mathbf{f}}(\mathbf{z_t}) &= \frac{1}{n}[(\sum_j \nabla f_j(\mathbf{z}_t^j))^\top \cdots (\sum_j \nabla f_j(\mathbf{z}_t^j))^\top]^\top,
\end{aligned} \tag{7}$$

and the global time-varying matrices

$$W^{(t)} = W_m^{(t)} \otimes I_d, \ \tilde{W}^{(t)} = \tilde{W}_m^{(t)} \otimes I_d. \tag{8}$$

The updates of global vectors $\mathbf{z}_t, \mathbf{h}_t \in \mathbb{R}^{nd}$ can be written as

$$\begin{aligned}
\mathbf{z}_{t+1} &= \tilde{W}^{(t)}(\mathbf{z}_t - \alpha \mathbf{h}_t), \\
\mathbf{h}_{t+1} &= \tilde{W}^{(t)}\mathbf{h}_t + \tilde{W}^{(t)}(Y_t^{-1} \otimes I_d)(\mathbf{v}_{t+1} - \mathbf{v}_t).
\end{aligned}$$

For the global vectors, let the $\mathfrak{L}$-norm be defined as $\mathfrak{L}^2(\mathbf{z}_t, \phi_t) = \|(diag(\phi_t)^{\frac{1}{2}} \otimes I_d)(\mathbf{z}_t - \hat{\mathbf{z}}_t)\|_F^2$, where $\hat{\mathbf{z}}_t$ is the ϕ_t-weighted average of $\mathbf{z}_t$. $\mathfrak{L}$-norm is induced by a sequence of time-varying stochastic vectors and facilitates the derivation of one-step consensus contraction addressed in Lemma 1 and 2.

Theorem 1. *Suppose Assumptions 1 – 3 hold. Let step size α satisfy*

$$\begin{aligned}
&0 < \alpha \leq \min\{\frac{(1-\delta^2)^2}{48\delta^2\|Y^{-1}\|\sqrt{(2L^2\phi_m(n+1) + 8L^4\phi_m(n+1))}}, \\
&\frac{1}{2L}, \frac{1-\delta^2}{8\delta\|Y^{-1}\|^{1/2}\sqrt[\frac{1}{4}]{(6L^2\phi_m(n+1) + L^2\phi_m)[\frac{18L^2}{(1-\delta^2)} + 36L^4]}}\}.
\end{aligned} \tag{9}$$

Moreover, let the momentum step size β be such that

$$48L^2\alpha^2 \leq \beta < 1. \tag{10}$$

Then it holds that

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\bar{\mathbf{x}}_t)\|^2 \\
&\leq \frac{2\mathbb{E}(f(\bar{\mathbf{x}}_0) - f(\bar{\mathbf{x}}_T))}{\alpha T} + \frac{2}{\beta T}\mathbb{E}[\|\bar{\mathbf{v}}_0 - \nabla \bar{\mathbf{f}}(\mathbf{z_0})\|^2] + 4\beta\bar{\nu}^2 \\
&+ \frac{16\delta^2\alpha^2}{(1-\delta^2)^4}(\frac{48L^2\phi_m(n+1)}{\beta n^2} + \frac{4L^2\phi_m(n+1)}{n}) \\
&\{\frac{(1-\delta^2)}{T}\mathbb{E}\mathfrak{L}^2(\mathbf{h}_0, \phi_0) + 48\delta^2\|Y^{-1}\|^2\beta^2 n\bar{\nu}^2 \\
&+ 48\delta^2\|Y^{-1}\|^2 L^2[\frac{\beta}{T}\mathbb{E}[\|\mathbf{v}_0 - \nabla \mathbf{f}(\mathbf{z_0})\|^2] + 2\beta^3 n\bar{\nu}^2]\},
\end{aligned} \tag{11}$$

where $\delta = \max_t \sqrt{1 - \frac{\min_i([\phi_{t+1}]_i)}{\max_i([\phi_t]_i)(n-1)^2 n^{2(n+2)}}} \in (0,1)$ is the network contraction parameter; $\phi_m = d/\min_{t,i}[\phi_t]_i$ is proportional to the inverse of the smallest entry in $\{\phi_t\}$, stochastic vectors associated with time-varying mixing matrices; and $\|Y^{-1}\| = \sup_t \|Y_t^{-1}\|$.[2]

Essentially, the bound in (11) is specified by the initial function value, the initial gradient estimation error, and the variance of the gradient estimator.

If step sizes α and β are chosen as in the statement of the above theorem, the average gradient error converges to a steady-state error,

$$\begin{aligned}
&\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\bar{\mathbf{x}}_t)\|^2 \\
&\leq \frac{2\mathbb{E}(f(\bar{\mathbf{x}}_0) - f(\bar{\mathbf{x}}_T))}{\alpha T} + \frac{2}{\beta T}\mathbb{E}[\|\bar{\mathbf{v}}_0 - \nabla \bar{\mathbf{f}}(\mathbf{z_0})\|^2] + 4\beta\bar{\nu}^2 \\
&+ \frac{768\delta^2\alpha^2}{(1-\delta^2)^4}(\frac{48L^2\phi_m(n+1)}{\beta n^2} + \frac{4L^2\phi_m(n+1)}{n}) \\
&(\delta^2\|Y^{-1}\|^2\beta^2 n\bar{\nu}^2 + 2\delta^2\|Y^{-1}\|^2 L^2\beta^3 n\bar{\nu}^2).
\end{aligned} \tag{12}$$

A closer inspection of the right-hand side reveals that by selecting appropriate α and β, one can provide non-asymptotic convergence guarantees. Appropriate choices of α and β and the corresponding convergence rate are specified in the following corollary.

Corollary 1.1. *There exist values of the parameters $\alpha = \mathcal{O}(\frac{1}{n^{1/2}T^{1/3}})$, $\beta = \mathcal{O}(\frac{1}{T^{2/3}})$ and $b = \mathcal{O}(\frac{T^{1/3}}{n})$ such that*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\bar{\mathbf{x}}_t)\|^2 \leq \mathcal{O}(\frac{1}{T^{2/3}}). \tag{13}$$

Recall that in each iteration of Push-ASGD, each node in the network samples one local data point and queries the stochastic first-order oracle for the gradient. Given the parameters in the corollary above, Push-ASGD can reach an ϵ-accurate stationary point of the global objective with the overall oracle complexity of $O(\frac{1}{\epsilon^{1.5}})$.

When comparing the oracle complexity of Push-ASGD with those of other SFO algorithms in Table I, we note that Push-ASGD has more desirable complexity for large T when n is fixed; although GT-HSGD has better oracle complexity, that algorithm does not apply to general directed time-varying graphs.

Note that Theorem 1 applies to any strongly connected network. In simulations we observe that the number of steps needed to reach desirable accuracy level is smaller in more densely connected networks because fewer gradient queries are required and more information is exchanged.

Next, we introduce the PL condition and investigate convergence under this additional assumption.

[1] Further details are in the proof of Lemma 1 in Appendix A.

[2] The existence/bounds for δ, ϕ_m, and $\|Y^{-1}\|$ are discussed in the proofs of Lemma 1 and 2.

This article has been accepted for publication in IEEE Transactions on Automatic Control. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAC.2024.3479888

Assumption 4. *The objective function satisfies the PL condition with parameter μ,*

$$\frac{1}{2}\|\nabla f(x)\|^2 \geq \mu(f(x) - f^*), \tag{14}$$

where f^* is the optimal value of the objective function.

Theorem 2. *Suppose Assumptions 1 – 4 hold. Let the step size α satisfy*

$$0 < \alpha \leq \min\{\frac{1}{2L}, \frac{1-\delta^2}{\mu}, \frac{(1-\delta^2)^4\mu}{36864\delta^4\|Y^{-1}\|^2L^2\phi_m(n+1)(3-\delta^2)}, \frac{1-\delta^2}{24\delta^2L\|Y^{-1}\|\sqrt{(72\phi_m(n+1)\frac{1}{(1-\delta^2)^2} + 16\phi_m(n+1)\frac{1}{1-\delta^2})}}\}. \tag{15}$$

Moreover, let the momentum step size β satisfy

$$\max\{\frac{\alpha\mu}{2}, \frac{768\alpha}{\mu}[\frac{3L^2\delta^4\phi_m(356+212n)\|Y^{-1}\|^2}{(1-\delta^2)^4} + \frac{3L^2}{2}] + 768\alpha^2[\frac{2736L^2\delta^4\phi_m(n+1)\|Y^{-1}\|^2}{(1-\delta^2)^4} + \frac{3L^2}{4(1-\delta^2)^2}]\} \leq \beta < 1. \tag{16}$$

Then $\mathbb{E}[f(\bar{\mathbf{x}}_{t+1}) - f^*]$ decays linearly at the rate of $\mathcal{O}((1-\frac{\alpha\mu}{4})^t)$ to a steady-state error, i.e.,

$$\begin{aligned}&\lim_{t\to\infty}\sup\mathbb{E}[f(\bar{\mathbf{x}}_{t+1}) - f^*]\\ &\leq (\frac{\alpha^2L^2}{4} + \frac{3L^2\alpha^2}{2\beta})(4\frac{C_{exp}}{\alpha} + 6\beta^2\bar{\nu}^2 + \frac{(1-\delta^2)^2}{\alpha}3\beta^2\bar{\nu}^2)\\ &+ \frac{288\beta\bar{\nu}^2}{\mu(1-\delta^2)^2}[\frac{8L^2\phi_m\alpha^2\delta^4(60+56\beta^2)(n+1)}{(1-\delta^2)^2}\|Y^{-1}\|^2\\ &+ (\frac{(1-\delta^2)^2}{4} + \frac{L^2\|Y^{-1}\|^2\alpha^2\delta^4\phi_m(96\beta^2+144\delta^2+96\beta^2n)}{(1-\delta^2)^2})],\end{aligned} \tag{17}$$

where $C_{exp} = \max_{k_m}\{(1-\frac{\alpha\mu}{2})[C^{k_m}\mathbf{u}_0]_4 + \frac{\alpha}{n}[C^{k_m}\mathbf{u}_0]_3 + \frac{\alpha L^2\phi_m(2n+2)}{n}[C^{k_m}\mathbf{u}_0]_1\}$ and $\mathbf{u}_0 = [\mathbb{E}\mathfrak{L}^2(\mathbf{z}_0,\phi_0), \mathbb{E}\mathfrak{L}^2(\mathbf{h}_0,\phi_0), \mathbb{E}[\|\mathbf{v}_0 - \nabla\mathbf{f}(\mathbf{z_0})\|_F^2], \mathbb{E}[f(\bar{\mathbf{x}}_0) - f^*]]$.

Theorem 2 implies that for small values of step sizes α and β (which satisfy the above conditions), the steady-state error will be small. Moreover, the following corollary on the non-asymptotic convergence holds.

Corollary 2.1. *Suppose Assumptions 1 - 4 and step size conditions (15) and (16) are satisfied. If the values of the step sizes are such that $\alpha = \mathcal{O}(T^{-1})$ and $\beta = o(T^{-1})$, i.e., $\alpha \to 0$ faster than $\beta \to 0$, then non-asymptotic convergence is guaranteed,*

$$\lim_{t\to\infty}\sup\mathbb{E}[f(\bar{\mathbf{x}}_{t+1}) - f^*] \to 0. \tag{18}$$

A. Sketch of the proof

Here we briefly go over the main steps of the proofs of Theorems 1-2; full details are presented in the appendix. First, we identify the main error terms that contribute to the overall convergence error of Push-ASGD. These include:

1. $\mathbb{E}[\mathfrak{L}^2(\mathbf{z}_t,\phi_t)]$: the consensus error that quantifies how far the local models are from their average formed via the weight matrix.
2. $\mathbb{E}\|\mathbf{v}_t - \nabla\mathbf{f}(\mathbf{z_t})\|^2$: error of the momentum-based stochastic gradient estimator.
3. $\mathbb{E}[\mathfrak{L}^2(\mathbf{h}_t,\phi_t)]$: error of the global gradient tracking estimator.
4. $\mathbb{E}\|\bar{\mathbf{v}}_t - \nabla\bar{\mathbf{f}}(\mathbf{z_t})\|^2$: variance of the momentum-based stochastic gradient estimator.
5. $\mathbb{E}[f(\bar{\mathbf{x}}_{t+1}) - f^*]$: optimality gap, measuring the distance from the optimal function value.

Our aim is to derive recursive inequalities that relate these error terms to each other. Then, to prove Theorem 1 we derive upper bounds on the relevant terms for each iteration t and sum them over t from 0 to T. Finally, we combine the intermediate steps to achieve the main result, i.e., establish a bound on the average gradient norm accumulated over the iterations. The major challenge in the analysis is to establish relationships between the following terms:

1. the consensus errors of the time-varying directed network system;
2. the combination of the global gradient tracking error originating due to communication of gradient information over the network, the stochastic gradient computation and the momentum term for convergence acceleration at local clients.

To start, we introduce a lemma specifying an upper bound on the consensus error at time t.

Lemma 1. *Suppose Assumptions 1 – 3 hold. Based on the updates of Push-ASGD,*

$$\mathbb{E}[\mathfrak{L}^2(\mathbf{z}_{t+1},\phi_{t+1})] \leq \frac{1+\delta^2}{2}\mathbb{E}\mathfrak{L}^2(\mathbf{z}_t,\phi_t) + \frac{2\delta^2\alpha^2}{1-\delta^2}\mathbb{E}\mathfrak{L}^2(\mathbf{h}_t,\phi_t), \tag{19}$$

for some network topology parameter $0 < \delta < 1$ indicated in Theorem 1.

Next, we present a lemma stating an upper bound on the gradient tracking error at time t.

Lemma 2. *Suppose Assumptions 1 – 3 hold. Then the gradient tracking error satisfies*

$$\begin{aligned}\mathbb{E}[\mathfrak{L}^2(\mathbf{h}_{t+1},\phi_{t+1})] \leq{}& \frac{1+\delta^2}{2}\mathbb{E}[\mathfrak{L}^2(\mathbf{h}_t,\phi_t)]\\ &+ \frac{8\delta^2}{1-\delta^2}\|Y^{-1}\|^2[3L^2[3\alpha^2\mathbb{E}[\|\bar{\mathbf{v}}_t\|^2]\\ &+ 6\phi_m(n+1)\mathbb{E}[\mathfrak{L}^2(\mathbf{z}_{t+1},\phi_{t+1}) + \mathfrak{L}^2(\mathbf{z}_t,\phi_t)]]\\ &+ 3\beta^2\mathbb{E}[\|\mathbf{v}_t - \nabla\mathbf{f}(\mathbf{z_t})\|^2] + 3\beta^2\bar{\nu}^2n].\end{aligned} \tag{20}$$

The following lemma states an upper bound on the error of the momentum-based stochastic gradient estimator.

Lemma 3. *Suppose Assumptions 1 – 3 hold. The error of the momentum-based stochastic gradient estimator satisfies*

$$\begin{aligned}&\mathbb{E}[\|\mathbf{v}_t - \nabla\mathbf{f}(\mathbf{z_t})\|^2] \leq (1-\beta)^2\mathbb{E}[\|\mathbf{v}_{t-1} - \nabla\mathbf{f}(\mathbf{z_{t-1}})\|^2] + 2\beta^2n\bar{\nu}^2\\ &+ 6(1-\beta)^2L^2[\alpha^2\mathbb{E}[\|\bar{\mathbf{v}}_{t-1}\|^2]\\ &+ 2\phi_m(n+1)\mathbb{E}[\mathfrak{L}^2(\mathbf{z}_t,\phi_t) + \mathfrak{L}^2(\mathbf{z}_{t-1},\phi_{t-1})\|^2]]\end{aligned} \tag{21}$$

while for its averaged version holds that

$$\begin{aligned}&\mathbb{E}[\|\bar{\mathbf{v}}_t - \nabla\bar{\mathbf{f}}(\mathbf{z_t})\|^2] \leq (1-\beta)^2\mathbb{E}[\|\bar{\mathbf{v}}_{t-1} - \nabla\bar{\mathbf{f}}(\mathbf{z_{t-1}})\|^2] + \frac{2\beta^2\bar{\nu}^2}{n}\\ &+ \frac{6(1-\beta)^2L^2}{n^2}[\alpha^2\mathbb{E}[\|\bar{\mathbf{v}}_{t-1}\|^2]\\ &+ 2\phi_m(n+1)\mathbb{E}[\mathfrak{L}^2(\mathbf{z}_t,\phi_t) + \mathfrak{L}^2(\mathbf{z}_{t-1},\phi_{t-1})\|^2]].\end{aligned} \tag{22}$$

Using the lemmas above, we proceed by conducting telescoping from $t = 0$ to T for each of the four errors to arrive at an upper bound for the sum of the errors. This leads to the following lemma for the squared gradient bound.

Lemma 4. *The accumulated expected gradient norm at $\bar{\mathbf{x}}_t$ satisfies*

$$\sum_{t=0}^{T-1} \mathbb{E}\|\nabla[f(\bar{\mathbf{x}}_t)\|^2] \leq \frac{2\mathbb{E}(f(\bar{\mathbf{x}}_0) - f(\bar{\mathbf{x}}_T))}{\alpha} - \frac{1}{2}\sum_{t=0}^{T-1} \mathbb{E}\|\bar{\mathbf{v}}_t\|^2$$
$$+ 2\sum_{t=0}^{T-1} \mathbb{E}\|\bar{\mathbf{v}}_t - \nabla\bar{\mathbf{f}}(\mathbf{z_t})\|^2 + \frac{4L^2\phi_m(n+1)}{n}\sum_{t=0}^{T-1} \mathbb{E}[\mathfrak{L}^2(\mathbf{z}_t, \phi_t)]. \tag{23}$$

To complete the proof of Theorem 1 we use the upper bounds on the last three sums on the right-hand-side of (23), canceling out the sum $\sum_{t=0}^{T-1} \mathbb{E}\|\bar{\mathbf{v}}_t\|^2$ which appears with a negative sign in (23). Doing so requires imposing limitations on α and β, the learning rates of Push-ASGD, as stated in Theorem 1.

Next, we outline the proof of Theorem 2, starting by building a linear inequality system on top of the error bound lemmas via incorporating the PL condition. In particular, one first needs to show that the largest eigenvalue of the coefficient matrix for the linear inequality system is strictly less than 1, guaranteeing linear convergence. Then, one needs to find an upper bound on the time-varying residual terms, including $\mathbb{E}\|\bar{\mathbf{v}}_t\|^2$, that holds for all t. The proof can finally be complected by imposing conditions on the step sizes α and β. The main challenges in this analysis stem from the following:

1) due to non-doubly-stochastic mixing matrices, the error terms in the linear inequality systems are considerably more involved than in the case of undirected static networks;
2) the dynamic residual term $\mathbf{r}_t$ contains $\mathbb{E}\|\bar{\mathbf{v}}_t\|^2$ and its upper bound.

To proceed, we use the PL condition to establish the following recursive relationship.

Lemma 5. *Suppose Assumptions 1 - 4 hold and the step size α is such that $\alpha \leq \frac{1}{2L}$. Then*

$$\mathbb{E}[f(\bar{\mathbf{x}}_{t+1}) - f^*] \leq \mathbb{E}[(1 - \frac{\alpha\mu}{2})(f(\bar{\mathbf{x}}_t) - f^*) - \frac{\alpha}{4}\|\bar{\mathbf{v}}_t\|^2$$
$$+ \frac{\alpha}{n}\|\mathbf{v}_t - \nabla\mathbf{f}(\mathbf{z_t})\|^2 + \frac{\alpha L^2\phi_m(2n+2)\mathfrak{L}^2(\mathbf{z}_t, \phi_t)}{n}]. \tag{24}$$

To form the linear inequality system, let us introduce

$$\mathbf{u}_{k+1} = \begin{bmatrix} \mathbb{E}[\mathfrak{L}^2(\mathbf{z}_{k+1}, \phi_{k+1})] \\ \mathbb{E}[\mathfrak{L}^2(\mathbf{h}_{k+1}, \phi_{k+1})] \\ \mathbb{E}[\|\mathbf{v}_{k+1} - \nabla\mathbf{f}(\mathbf{z_{k+1}})\|_F^2] \\ \mathbb{E}[f(\bar{\mathbf{x}}_{k+1}) - f^*] \end{bmatrix},$$

$$C_1 = [\frac{1+\delta^2}{2}, \frac{2\delta^2\alpha^2}{1-\delta^2}, 0, 0]$$
$$C_2 = [\frac{144\delta^2\|Y^{-1}\|^2L^2\phi_m(2n+2)}{1-\delta^2}, \frac{1+\delta^2}{2}$$
$$+ \frac{288\delta^4\|Y^{-1}\|^2L^2\phi_m\alpha^2(n+1)}{1-\delta^2}, \frac{24\delta^2\beta^2\|Y^{-1}\|^2}{1-\delta^2}, 0]$$
$$C_3 = [24(1-\beta)^2L^2\phi_m(n+1), 24(1-\beta)^2L^2\phi_m(n+1)\frac{\delta^2\alpha^2}{1-\delta^2},$$
$$(1-\beta)^2, 0]$$
$$C_4 = [\frac{2\alpha L^2\phi_m(n+1)}{n}, 0, \frac{\alpha}{n}, 1 - \frac{\alpha\mu}{2}], \tag{25}$$

and

$$\mathbf{r}_k = \begin{bmatrix} 0 \\ \frac{8\delta^2}{1-\delta^2}\|Y^{-1}\|^2(9\alpha^2L^2\mathbb{E}[\|\bar{\mathbf{v}}_k\|^2] + 3\beta^2\bar{\nu}^2 n) \\ 6(1-\beta)^2L^2\alpha^2\mathbb{E}[\|\bar{\mathbf{v}}_k\|^2] + 2\beta^2\bar{\nu}^2 n \\ 0 \end{bmatrix}.$$

Moreover, let us for convenience denote

$$C = [C_1^T, C_2^T, C_3^T, C_4^T]^T. \tag{26}$$

It is straighforward to show that

$$\mathbf{u}_{k+1} \leq C\mathbf{u}_k + \mathbf{r}_k.$$

Therefore, if one could find $\mathbf{x} > 0$ such that $C\mathbf{x} < \mathbf{x}$, then $\rho(C) < 1$. The following lemma provides such a guarantee.

Lemma 6. *For the range of α in Lemma 5, one can find $\mathbf{x} > 0$ such that $\rho(C) \leq 1 - \frac{\alpha\mu}{4}$.*

It then follows that for $\forall t \in [1, T]$,

$$\mathbf{u}_t \leq C^t\mathbf{u}_0 + \sum_{k=0}^{t-1} C^{t-1-k}\mathbf{r}_k. \tag{27}$$

In the final stage of the argument, we establish an upper bound on $\mathbf{r}_k$ by bounding $E[\|\bar{\mathbf{v}}_k\|^2]$, and show that the bound is independent of the iteration index.

Lemma 7. *Suppose Assumptions 1 - 4 hold, and let α and β satisfy the conditions of Theorem 2. Then $E[\|\bar{\mathbf{v}}_t\|^2]$ can be upper bounded by a function of $\bar{\nu}$ which is independent of the iteration index t.*

The proof of Theorem 2 is completed by incorporating the above two lemmas into Lemma 5 and summarizing the required conditions for the step sizes.

V. Experimental Results

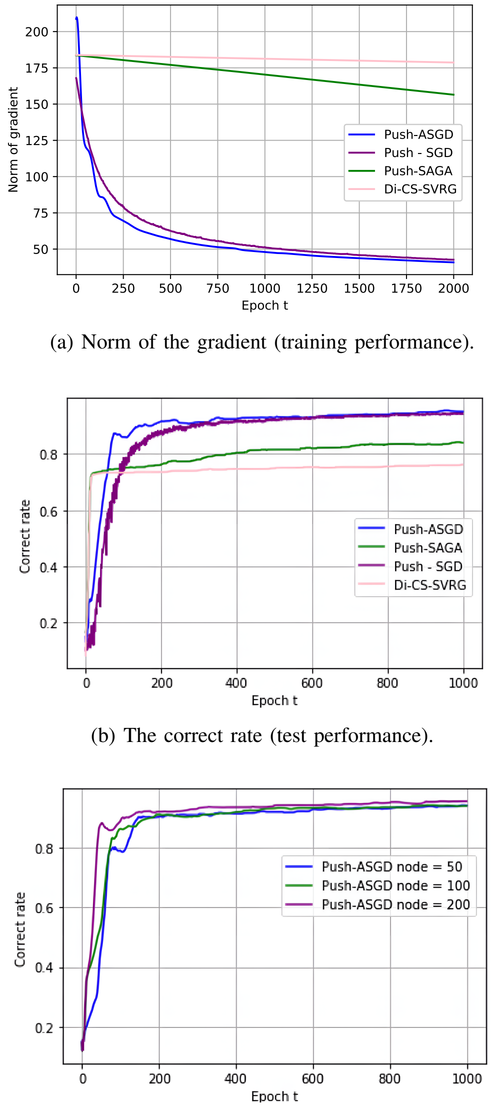
In this section we report performance of the proposed algorithm, Push-ASGD, in a variety of experimental settings, including three different ML tasks and a numerical study illustrating performance under PL condition. Specifically, Push-ASGD is benchmarked against the following methods tailored to directed networks: SGP/Push-SGD [29, 5], the Push-SAGA algorithm [12] and the Di-CS-SVRG algorithm [11]. The first of these algorithm uses only local stochastic gradient updates while the latter two incorporate both global gradient tracking and variance reduction. Please note that there are no theoretical guarantees for the latter two algorithms in decentralized non-convex settings; moreover, those two schemes assume access to IFO while our Push-ASGD is tailored to the more challenging SFO scenario. For all algorithms, the step sizes are selected from $[10^{-7}, 10^{-1}]$ while for Push-ASGD β is selected from $[10^{-4}, 10^{-1}]$; the best performance for each method is reported.

Regarding the experiments: we first perform a test on the logistic regression model with non-convex regularization; the second experiment is an image classification task on the CIFAR-10 dataset via a shallow neural network; the third experiment is an NLP classification task on the IMDB dataset via fine-tuning the BERT architecture. We further consider an objective function satisfying the PL condition and numerically test the algorithms' performance.

In all tasks, the training data is distributed randomly without shuffling; hence, nodes are not guaranteed to receive data from all classes. In other words, the nodes experience different data distributions due to a high degree of heterogeneity.

A. Non-convex logistic regression

We first benchmark the performance of Push-ASGD on a decentralized non-convex logistic regression model applied to handwritten digit classification task on the MNIST dataset [32]. The learning task is distributed across 100 nodes of a time-varying network generated according to the Erdős-Rényi model and as a directed ring. For the Erdős-Rényi model, we generate the graph and randomly remove a subset of edges to make the graph directed. The network is switching



(a) Norm of the gradient (training performance).

(b) The correct rate (test performance).

(c) Performance as the network size varies.

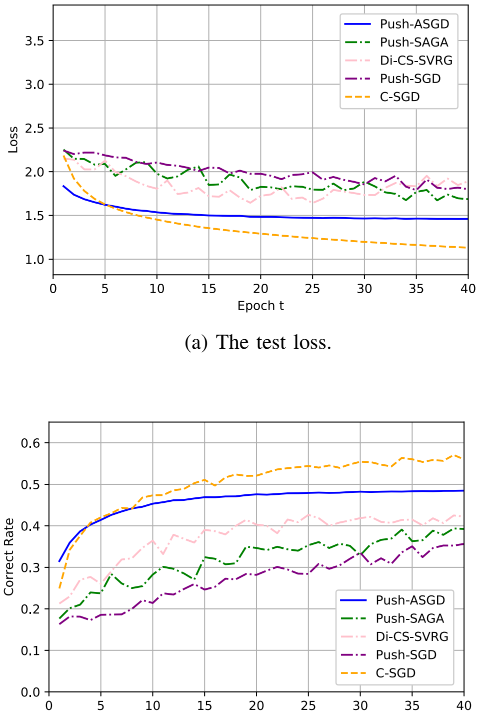
Fig. 1. Performance on MNIST. Push-ASGD achieves lower loss and higher correct rate than the competing schemes.

between the Erdős-Rényi model, a directed ring and a reversed directed ring. The dataset is distributed such that each node has 12 images for local training. We deploy a non-convex regularizer and consider the minimization [33]

$$\min_{\mathbf{x}} \left\{ \sum_{i=1}^{n} \sum_{j=1}^{N} \ln(1 + e^{-(\mathbf{m}_{ij}^T \mathbf{x})\mathbf{y}_{ij}}) + \sum_{j=1}^{d} \frac{\lambda [\mathbf{x}]_j^2}{1 + [\mathbf{x}]_j^2} \right\}, \tag{28}$$

where $(\mathbf{m}_{ij}, \mathbf{y}_{ij})$ represents the image feature vector and the corresponding label of the j-th image at node i. Parameters of the algorithms are set to $\alpha = 6 \times 10^{-5}$ and $\beta = 0.015$ (Push-ASGD), $\alpha = 6 \times 10^{-5}$ (Subgradient-Push with SGD), and $\alpha = 2 \times 10^{-5}$ and 2×10^{-7} (Di-CS-SVRG and Push-SAGA, respectively). The regularization parameter is set to $\lambda = 10^{-4}$. For all the experiments in this section, the batch size is set to 1.

Results of the benchmarking experiments on the non-convex logistic regression task are shown in Figure 1. As can be seen there, the accuracy achieved by Push-ASGD is the highest among all the considered schemes. This confirms our expectation that Push-SGD should outperform Push-SAGA and Di-CS-SVRG since the accuracy of the latter two schemes is adversely affected by the sparse connectivity of the considered directed ring graph structure (such a structure causes instability of primal-dual schemes [34, 35]).[3]



(a) The test loss.

(b) The correct rate.

Fig. 2. Performance on CIFAR-10. Push-ASGD achieves lower loss and higher correct rate than the competing schemes.

We also test the performance of Push-ASGD as the network size varies; in particular, we increase the number of network nodes from 50 to 200. For fixed network connectivity level and number of local data points, the convergence is faster when there are more nodes in the network (see Figure 1(c)).

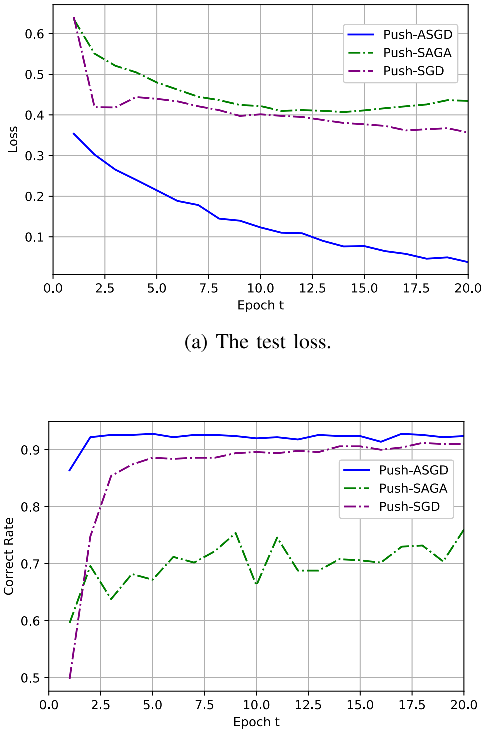
B. Image classification experiments

Next, we test the performance of the proposed algorithm on a decentralized image classification task involving CIFAR-10 dataset [36]. To this end, we rely on a convolutional neural network architecture Lenet [32]. Lenet consists of 5 layers: two sets of convolutional, activation, and max-pooling layers, followed by two fully-connected layers with activation and a softmax classifier. For this task we utilize the cross-entropy loss.

The time-varying directed network is constructed based on the Erdős-Rényi model and directed rings. First, the Erdős-Rényi graph with 10 nodes is generated, and then several edges are removed to induce a directed graph. Each node is assigned 5000 images from the CIFAR-10 dataset for local training. For all algorithms, step size is set to $\alpha = 10^{-2}$; for Push-ASGD, the momentum step size is set to $\beta = 0.05$. As a reference, we also provide a comparison with the centralized stochastic gradient descent (C-SGD), for which the training data includes all 50000 images. For all the experiments in this section, the batch size is 32.

The test loss and the correct rate are reported in Figures 2(a) and 2(b), respectively. As seen there, the proposed algorithm, Push-ASGD,

[3] A similar phenomenon is observed in Fig. 3; in Fig. 2, Push-SAGA and Di-CS-SVRG outperform Push-SGD due to a high number of local training data points and reduced stochasticity.



(a) The test loss.

(b) The correct rate.

Fig. 3. Performance on the natural language processing task. Push-ASGD achieves lower loss and higher accuracy than the competing schemes.

outperforms other decentralized schemes. The gap between Push-ASGD and C-SGD is due to the impact of distributing the dataset across the network nodes while maintaining the same total amount of data as used by the centralized method.

C. Natural language processing experiments

The remaining real-world data experiment involves an NLP classification task via fine-tuning a deep learning language model. In particular, we train this model on the IMDB dataset that contains the texts of reviews and the corresponding binary tags implying whether the review is positive or negative [37]. We still consider an Erdős-Rényi-based directed time-varying network and distribute the IMDB training data such that each node has access to 2000 reviews and uses them for local training. The model is constructed by adding a linear classification layer to the pre-trained Bidirectional Encoder Representations from Transformers (BERT) [38] architecture, and fine-tuned locally. We again utilize the cross-entropy loss.

The Push-ASGD and Push-SGD algorithms use step size $\alpha = 0.002$, while Push-SAGA uses a smaller step size $\alpha = 0.0002$ to avoid divergence. For the Push-ASGD algorithm, the momentum step size is set to $\beta = 0.025$. In all the experiments in this section the batch size is set to 4. The performance of the algorithms is shown in Fig. 3. As seen there, Push-ASGD achieves lower loss and converges to the highest correct rate of over 90%.[4]

D. The PL condition

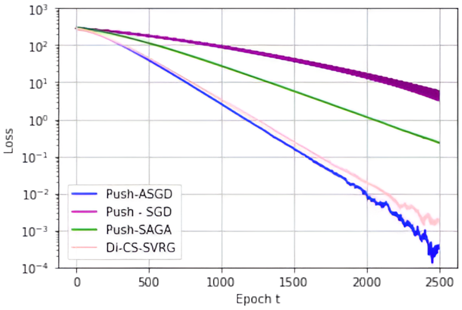


Fig. 4. In simulations of a setting where PL condition holds, Push-ASGD converges faster than other benchmarking algorithms.

Lastly, we test the performance of the algorithms in an application to minimizing a global function that satisfies the PL condition. The local functions is defined as $f_i(x) = x^2 + 3\sin^2(x) + a_i \cos(x)$, where a_i are non-zero parameters satisfying $\sum_{i=1}^{n} a_i = 0$ so that the global function is $F(x) = x^2 + 3\sin^2(x)$. The global function is non-convex and satisfies the PL condition [39]. To simulate stochastic gradient, we add random Gaussian noise with mean 0 and standard deviation $1/2$ to the gradient at each node. The network consists of 100 nodes; similar to the previous experiments, the time-varying directed network topology is based on the Erdős-Rényi model and directed rings. As seen in Fig. 4, Push-ASGD converges the fastest, followed by the two benchmarking algorithms with gradient acceleration; Push-SGD, computing a simple stochastic gradient, converges the slowest and is less stable than other algorithms.

VI. Conclusion

The paper presents the first analytical study of decentralized stochastic non-convex optimization over time-varying directed graphs, and introduces a novel stochastic optimization algorithm for this problem. The method, Push-ASGD, is the first scheme that achieves the SFO complexity of $\mathcal{O}(1/\epsilon^{1.5})$ for smooth objectives; in addition, it enjoys linear convergence under the PL condition. Push-ASGD relies on a push-sum protocol to perform local aggregation under communication asymmetry, while employing a novel stochastic gradient estimator to deal with uncertainties stemming from noise and heterogeneity in local data. The proposed gradient estimator incorporates momentum-based variance reduction and gradient tracking techniques to recursively estimate global gradient, which is unknown to the participating agents. Extensive experiments demonstrate that Push-ASGD outperforms existing methods for distributed optimization over time-varying directed networks.

VII. Appendix

In the appendix we provide details of the analysis summarized in the theorems.

A. Proof of Lemma 1

We start by defining the norm with respect to time-varying stochastic vectors, ϕ_t. Recall the update of $\mathbf{z}_t$,

$$\mathbf{z}_{t+1} = \tilde{W}^{(t)}(\mathbf{z}_t - \alpha \mathbf{h}_t),$$

where $\tilde{W}^{(t)} = \tilde{W}_m^{(t)} \otimes I_d$ and $\mathbf{z}_t, \mathbf{h}_t \in \mathbb{R}^{nd}$. The (i,j)-th entry in the column-stochastic matrix $W_m^{(t)}$ is given by

$$w_{ij}^{(t)} = \frac{1}{d_j^{out,t} + 1} \quad \text{for} \quad (j,i) \in \mathcal{E}(t), \tag{29}$$

where $d_j^{out,t}$ denotes the out-degree of agent j at time t. Since the entries of $\tilde{W}_m^{(t)}$ are given by $\tilde{w}_{i,j}^{(t)} = w_{i,j}^{(t)} y_t^j / y_{t+1}^i$, it is guaranteed that each row of $\tilde{W}_m^{(t)}$ has row sum equal to 1 [10]. Since $\tilde{W}_m^{(t)}$ is row-stochastic and the network is strongly connected, $\tilde{W}_m^{(t)}$ can also be viewed as a row-stochastic mixing matrix for strongly-connected

[4] The performance of Di-CS-SVRG is omitted since it failed to converge in 20 epochs.

directed graphs. Using the weight policy of W_m^t and the update of y_t, we can derive that the positive entries in $\tilde{W}_m^{(t)}$ can be uniformly lower bounded as $\tilde{w}_{i,j}^{(t)} \geq \omega = \frac{1}{n^{n+2}}$ for $(j,i) \in \mathcal{E}(t)$ [10]. Consider the stochastic vector sequence $\{\phi_t\}$ such that ${\phi_{t+1}}^T \tilde{W}_m^{(t)} = \phi_t^T$; from Lemma 3.3 of [25], the sequence $\{\phi_t\}$ exists and has element-wise lower bound, i.e., $[\phi_t]_i \geq \frac{\omega^n}{n}$. We recall the definition of $\mathfrak{L}^2(\mathbf{z}_t, \phi_t)$,

$$\mathfrak{L}^2(\mathbf{z}_t, \phi_t) = \|(diag(\phi_t)^{\frac{1}{2}} \otimes I_d)(\mathbf{z}_t - \hat{\mathbf{z}}_t)\|_F^2,$$

where $\hat{\mathbf{z}}_t$ is the ϕ_t-weighted average of $\mathbf{z}_t$ in $\mathcal{R}^{nd}$. Following Lemma 4.2 in [25], we can obtain $\mathfrak{L}(\tilde{W}_m^{(t)}\mathbf{z}, \phi_{t+1}) \leq \lambda_t \mathfrak{L}(\mathbf{z}, \phi_t)$, where $\lambda_t = \sqrt{1 - \frac{\min_i([\phi_{t+1}]_i)\omega^2}{\max_i([\phi_t]_i)(n-1)^2}} \in (0,1)$. Using the triangle and Young's inequalities we obtain

$$\begin{aligned}
&\mathfrak{L}^2(\mathbf{z}_{t+1}, \phi_{t+1}) = \mathfrak{L}^2(\tilde{W}^{(t)}(\mathbf{z}_t - \alpha \mathbf{h}_t), \phi_{t+1}) \\
&\leq (1+r)\mathfrak{L}^2(\tilde{W}^{(t)}\mathbf{z}_t, \phi_{t+1}) + (1+\frac{1}{r})\alpha^2 \mathfrak{L}^2(\tilde{W}^{(t)}\mathbf{h}_t, \phi_{t+1}) \\
&\leq (1+r)\delta^2 \mathfrak{L}^2(\mathbf{z}_t, \phi_t) + (1+\frac{1}{r})\delta^2\alpha^2 \mathfrak{L}^2(\mathbf{h}_t, \phi_t) \\
&\leq \frac{1+\delta^2}{2}\mathfrak{L}^2(\mathbf{z}_t, \phi_t) + \frac{(1+\delta^2)\delta^2\alpha^2}{1-\delta^2}\mathfrak{L}^2(\mathbf{h}_t, \phi_t) \\
&\leq \frac{1+\delta^2}{2}\mathfrak{L}^2(\mathbf{z}_t, \phi_t) + \frac{2\delta^2\alpha^2}{1-\delta^2}\mathfrak{L}^2(\mathbf{h}_t, \phi_t),
\end{aligned} \tag{30}$$

where $r = \frac{1-\delta^2}{2\delta^2}$ and $\delta = \max_{t \geq 0} \lambda_t$. The proof of Lemma 1 is completed by taking the expectation of both sides of the inequality.

B. Proof of Lemma 2

We start by applying techniques similar to those used in the consensus error analysis in the proof of Lemma 1, yielding

$$\begin{aligned}
&\mathfrak{L}^2(\mathbf{h}_{t+1}, \phi_{t+1}) \\
&= \mathfrak{L}^2(\tilde{W}^{(t)}\mathbf{h}_t + \tilde{W}^{(t)}(Y_t^{-1} \otimes I_d)(\mathbf{v}_{t+1} - \mathbf{v}_t), \phi_{t+1}) \\
&\leq \frac{1+\delta^2}{2}\mathfrak{L}^2(\mathbf{h}_t, \phi_t) + \frac{2\delta^2}{1-\delta^2}\|Y^{-1}\|^2 \mathfrak{L}^2(\mathbf{v}_{t+1} - \mathbf{v}_t, \phi_{t+1}) \\
&\leq \frac{1+\delta^2}{2}\mathfrak{L}^2(\mathbf{h}_t, \phi_t) + \frac{8\delta^2}{1-\delta^2}\|Y^{-1}\|^2 \|\mathbf{v}_{t+1} - \mathbf{v}_t\|_2^2
\end{aligned}$$

where $\|Y^{-1}\| = \sup_t \|Y_t^{-1}\|_{\max}$ denotes the supremum of the inverse elements across all iterations, and the second inequality follows from the definition of the $\mathcal{L}$-norm. Moreover, $\|Y^{-1}\| \leq n^n$ since the smallest element of $\|Y_t\|_{\max} \geq \frac{1}{n^n}$ for any t [10]. After taking expectation of both sides, we obtain

$$\begin{aligned}
\mathbb{E}\mathfrak{L}^2(\mathbf{h}_{t+1}, \phi_{t+1}) \quad &\leq \quad \frac{1+\delta^2}{2}\mathbb{E}\mathfrak{L}^2(\mathbf{h}_t, \phi_t) \\
&+ \quad \frac{8\delta^2\|Y^{-1}\|^2}{1-\delta^2}\mathbb{E}\|\mathbf{v}_{t+1} - \mathbf{v}_t\|_2^2.
\end{aligned} \tag{31}$$

We proceed by deriving an upper bound on $\mathbb{E}\|\mathbf{v}_{t+1} - \mathbf{v}_t\|_2^2$. To this end, note that for each i,

$$\begin{aligned}
\mathbf{v}_{t+1}^i - \mathbf{v}_t^i &= \nabla f_i(\mathbf{z}_{t+1}^i, \xi_t^i) + (1-\beta)(\mathbf{v}_t^i - \nabla f_i(\mathbf{z}_t^i, \xi_t^i)) - \mathbf{v}_t^i \\
&= \nabla f_i(\mathbf{z}_{t+1}^i, \xi_t^i) - \nabla f_i(\mathbf{z}_t^i, \xi_t^i) - \beta \mathbf{v}_t^i + \beta \nabla f_i(\mathbf{z}_t^i, \xi_t^i) \\
&= \nabla f_i(\mathbf{z}_{t+1}^i, \xi_t^i) - \nabla f_i(\mathbf{z}_t^i, \xi_t^i) \\
&- \beta(\mathbf{v}_t^i - \nabla \mathbf{f}_i(\mathbf{z_t^i})) + \beta(\nabla f_i(\mathbf{z}_t^i, \xi_t^i) - \nabla \mathbf{f}_i(\mathbf{z_t^i})).
\end{aligned}$$

Taking the expectation of $\|\mathbf{v}_{t+1}^i - \mathbf{v}_t^i\|^2$,

$$\begin{aligned}
&\mathbb{E}[\|\mathbf{v}_{t+1}^i - \mathbf{v}_t^i\|^2] \\
&\overset{(a)}{\leq} 3\mathbb{E}[\|\nabla f_i(\mathbf{z}_{t+1}^i, \xi_t^i) - \nabla f_i(\mathbf{z}_t^i, \xi_t^i)\|^2] + \\
&\quad 3\beta^2\mathbb{E}[\|\mathbf{v}_t^i - \nabla \mathbf{f}_i(\mathbf{z_t^i})\|^2] + 3\beta^2\mathbb{E}[\|\nabla f_i(\mathbf{z}_t^i, \xi_t^i) - \nabla \mathbf{f}_i(\mathbf{z_t^i})\|^2] \\
&\overset{(b)}{\leq} 3L^2\mathbb{E}[\|\mathbf{z}_{t+1}^i - \mathbf{z}_t^i\|^2] + 3\beta^2\mathbb{E}[\|\mathbf{v}_t^i - \nabla \mathbf{f}_i(\mathbf{z_t^i})\|^2] + 3\beta^2\nu_i^2,
\end{aligned}$$

where (a) is due to Cauchy–Schwarz inequality while for (b) we invoke the smoothness assumption; here L denotes the smoothness parameter and ν_i^2 is a bound on the stochastic gradient estimate for given i. Summing up from $i = 1$ to n yields

$$\begin{aligned}
&\mathbb{E}[\|\mathbf{v}_{t+1} - \mathbf{v}_t\|^2] \\
&\overset{(c)}{\leq} 3L^2\mathbb{E}[\|\mathbf{z}_{t+1} - \mathbf{z}_t\|^2] + 3\beta^2\mathbb{E}[\|\mathbf{v}_t - \nabla \mathbf{f}(\mathbf{z_t})\|^2] + 3\beta^2\bar{\nu}^2 n \\
&= 3L^2\mathbb{E}[\|\mathbf{z}_{t+1} - \bar{\mathbf{x}}_{t+1} + \bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_t + \bar{\mathbf{x}}_t - \mathbf{z}_t\|^2] \\
&+ 3\beta^2\mathbb{E}[\|\mathbf{v}_t - \nabla \mathbf{f}(\mathbf{z_t})\|^2] + 3\beta^2\bar{\nu}^2 n \\
&\overset{(d)}{\leq} 3L^2[3\mathbb{E}[\|\bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_t\|^2] + 3\mathbb{E}[\|\mathbf{z}_{t+1} - \bar{\mathbf{x}}_{t+1}\|^2 + \|\mathbf{z}_t - \bar{\mathbf{x}}_t\|^2]] \\
&+ 3\beta^2\mathbb{E}[\|\mathbf{v}_t - \nabla \mathbf{f}(\mathbf{z_t})\|^2] + 3\beta^2\bar{\nu}^2 n \\
&\overset{(e)}{\leq} 3L^2[3\alpha^2\mathbb{E}[\|\bar{\mathbf{v}}_t\|^2] + 3\mathbb{E}[\|\mathbf{z}_{t+1} - \bar{\mathbf{x}}_{t+1}\|^2 + \|\mathbf{z}_t - \bar{\mathbf{x}}_t\|^2]] \\
&+ 3\beta^2\mathbb{E}[\|\mathbf{v}_t - \nabla \mathbf{f}(\mathbf{z_t})\|^2] + 3\beta^2\bar{\nu}^2 n,
\end{aligned}$$

where $\bar{\mathbf{v}}_t = [(\frac{1}{n}\sum_j \mathbf{v}_t^j)^T, \cdots, (\frac{1}{n}\sum_j \mathbf{v}_t^j)^T]^T \in \mathbb{R}^{np}$ and $\bar{\nu}^2 = \frac{1}{n}\sum_{i=1}^n \nu_i^2$; note that (c) follows (b), (d) is due to the Cauchy-Schwarz inequality and (e) stems from the update rule of $\mathbf{x}_t$. The term $\mathbb{E}\|\mathbf{z}_{t+1} - \bar{\mathbf{x}}_{t+1}\|^2$ appears in the upper bound and thus deserves closer attention. Recall that $\hat{\mathbf{z}}_t$ is the ϕ_t-weighted average of $\mathbf{z}_t$; letting $\phi_m = d/\min_{t,i}[\phi_t]_i$, it holds that

$$\begin{aligned}
&\|\mathbf{z}_t - \bar{\mathbf{x}}_t\|^2 = \|\mathbf{z}_t - \hat{\mathbf{z}}_t + \hat{\mathbf{z}}_t - \bar{\mathbf{x}}_t\|_F^2 \\
&\overset{(f)}{\leq} 2\frac{1}{\min(\phi_t)}\mathfrak{L}^2(\mathbf{z}_t, \phi_t) + 2n\|\sum_{j=1}^n [\phi_t]_j \mathbf{z}_t^j - \frac{1}{n}\sum_{j=1}^n [y_t]_j \mathbf{z}_t^j\|_2^2 \\
&= 2\frac{1}{\min(\phi_t)}\mathfrak{L}^2(\mathbf{z}_t, \phi_t) + 2n\sum_{m=1}^d [\sum_{j=1}^n [\phi_t]_j \mathbf{z}_t^j - \frac{1}{n}\sum_{j=1}^n [y_t]_j \mathbf{z}_t^j]_m^2 \\
&\overset{(g)}{\leq} 2\frac{1}{\min(\phi_t)}\mathfrak{L}^2(\mathbf{z}_t, \phi_t) + 2n\sum_{m=1}^d \max_i [\sum_{j=1}^n [\phi_t]_j \mathbf{z}_t^j - \mathbf{z}_t^i]_m^2 \\
&\overset{(h)}{\leq} 2\frac{1}{\min(\phi_t)}\mathfrak{L}^2(\mathbf{z}_t, \phi_t) + 2n\sum_{m=1}^d \max_i \|\sum_{j=1}^n [\phi_t]_j \mathbf{z}_t^j - \mathbf{z}_t^i\|_2^2 \\
&= 2\frac{1}{\min(\phi_t)}\mathfrak{L}^2(\mathbf{z}_t, \phi_t) + 2nd\max_i \|\sum_{j=1}^n [\phi_t]_j \mathbf{z}_t^j - \mathbf{z}_t^i\|_2^2 \\
&\overset{(i)}{\leq} 2\frac{1}{\min(\phi_t)}\mathfrak{L}^2(\mathbf{z}_t, \phi_t) + 2nd\|\mathbf{z}_t - \hat{\mathbf{z}}_t\|_2^2 \\
&\overset{(j)}{\leq} 2\phi_m \mathfrak{L}^2(\mathbf{z}_t, \phi_t) + 2n\phi_m \mathfrak{L}^2(\mathbf{z}_t, \phi_t),
\end{aligned} \tag{32}$$

where (f) is due to the Cauchy-Schwarz inequality and the fact that column-stochasticity of matrix $W_m^{(t)}$ ensures that $\sum_{j=1}^n [y_t]_j = \sum_{j=1}^n [y_0]_j = n$. One can show (g) by first finding the maximum and minimum of the weighted sum $\sum_{i=1}^n a_i [\mathbf{z}_t^i]_m$ subject to $\sum_{i=1}^n a_i = 1, a_i \geq 0$. Specifically, the maximum is achieved by putting all the weight on the largest $[\mathbf{z}_t^i]_m$, while the minimum is achieved by putting all the weight on the smallest $[\mathbf{z}_t^i]_m$. Let $i^* = \text{argmax}_i [\mathbf{z}_t^i]_m$ and $j^* = \text{argmin}_i [\mathbf{z}_t^i]_m$; then $\max_{\{a_i\}}[\sum_{j=1}^n [\phi_t]_j \mathbf{z}_t^j - \sum_{j=1}^n a_j \mathbf{z}_t^j]_m^2 = \max\{[\sum_{j=1}^n [\phi_t]_j \mathbf{z}_t^j - \mathbf{z}_t^{i^*}]_m^2, [\sum_{j=1}^n [\phi_t]_j \mathbf{z}_t^j - \mathbf{z}_t^{j^*}]_m^2\}$. Therefore, it must be that $[\sum_{j=1}^n [\phi_t]_j \mathbf{z}_t^j - \sum_{j=1}^n a_j \mathbf{z}_t^j]_m^2 \leq \max_i [\sum_{j=1}^n [\phi_t]_j \mathbf{z}_t^j - \mathbf{z}_t^i]_m^2$. (h) is due to the fact that the square of each entry of a vector is no greater than the squared ℓ_2 norm of the vector. Finally, $(i) - (j)$ are due to the definition of $\mathfrak{L}^2(\mathbf{z}_t, \phi_t)$. Taking the expectation and revisiting the bound on $\mathbb{E}\|\mathbf{v}_{t+1}-\mathbf{v}_t\|_2^2$ yields

$$\begin{aligned}
&\mathbb{E}[\|\mathbf{v}_{t+1}-\mathbf{v}_t\|^2]\\
&=3L^2[3\alpha^2\mathbb{E}[\|\bar{\mathbf{v}}_t\|^2]+3\mathbb{E}[\|\mathbf{z}_{t+1}-\bar{\mathbf{x}}_{t+1}\|^2+\|\mathbf{z}_t-\bar{\mathbf{x}}_t\|^2]]\\
&+3\beta^2\mathbb{E}[\|\mathbf{v}_t-\nabla\mathbf{f}(\mathbf{z_t})\|^2]+3\beta^2\bar{\nu}^2n\\
&\leq 3L^2[3\alpha^2\mathbb{E}[\|\bar{\mathbf{v}}_t\|^2]\\
&+6\phi_m(n+1)\mathbb{E}[\mathfrak{L}^2(\mathbf{z}_{t+1},\phi_{t+1})+\mathfrak{L}^2(\mathbf{z}_t,\phi_t)]]\\
&+3\beta^2\mathbb{E}[\|\mathbf{v}_t-\nabla\mathbf{f}(\mathbf{z_t})\|^2]+3\beta^2\bar{\nu}^2n,
\end{aligned}$$

where the last inequality is due to (e) and (i). Substituting the bound on $\mathbb{E}[\|\mathbf{v}_{t+1}-\mathbf{v}_t\|^2]$ in (31), we obtain

$$\begin{aligned}
&\mathbb{E}[\mathfrak{L}^2(\mathbf{h}_{t+1},\phi_{t+1})]\\
&\leq\frac{1+\delta^2}{2}\mathbb{E}[\mathfrak{L}^2(\mathbf{h}_t,\phi_t)]+\frac{8\delta^2}{1-\delta^2}\|Y^{-1}\|^2\mathbb{E}\|\mathbf{v}_{t+1}-\mathbf{v}_t\|^2]\\
&\leq\frac{1+\delta^2}{2}\mathbb{E}[\mathfrak{L}^2(\mathbf{h}_t,\phi_t)]+\frac{8\delta^2}{1-\delta^2}\|Y^{-1}\|^2[3L^2[3\alpha^2\mathbb{E}[\|\bar{\mathbf{v}}_t\|^2]\\
&+6\phi_m(n+1)\mathbb{E}[\mathfrak{L}^2(\mathbf{z}_{t+1},\phi_{t+1})+\mathfrak{L}^2(\mathbf{z}_t,\phi_t)]]\\
&+3\beta^2\mathbb{E}[\|\mathbf{v}_t-\nabla\mathbf{f}(\mathbf{z_t})\|^2]+3\beta^2\bar{\nu}^2n].
\end{aligned}$$

C. Proof of Lemma 3

Recall the rule for updating the local gradient estimate $\mathbf{v}_t^i$,

$$\begin{aligned}
\mathbf{v}_t^i &= \nabla f_i(\mathbf{z}_t^i,\xi_{t-1}^i)+(1-\beta)(\mathbf{v}_{t-1}^i-\nabla f_i(\mathbf{z}_{t-1}^i,\xi_{t-1}^i))\\
&= \beta\nabla f_i(\mathbf{z}_t^i,\xi_{t-1}^i)\\
&+ (1-\beta)(\mathbf{v}_{t-1}^i+\nabla f_i(\mathbf{z}_t^i,\xi_{t-1}^i)-\nabla f_i(\mathbf{z}_{t-1}^i,\xi_{t-1}^i)).
\end{aligned}$$

For all $t\geq 1$ and i,

$$\begin{aligned}
&\mathbf{v}_t^i-\nabla\mathbf{f}_i(\mathbf{z_t^i})=\beta\nabla f_i(\mathbf{z}_t^i,\xi_{t-1}^i)+(1-\beta)(\mathbf{v}_{t-1}^i+\nabla f_i(\mathbf{z}_t^i,\xi_{t-1}^i)\\
&-\nabla f_i(\mathbf{z}_{t-1}^i,\xi_{t-1}^i))-\beta\nabla\mathbf{f}_i(\mathbf{z_t^i})-(1-\beta)\nabla\mathbf{f}_i(\mathbf{z_t^i})\\
&=\beta(\nabla f_i(\mathbf{z}_t^i,\xi_{t-1}^i)-\nabla\mathbf{f}_i(\mathbf{z_t^i}))+(1-\beta)(\mathbf{v}_{t-1}^i+\nabla f_i(\mathbf{z}_t^i,\xi_{t-1}^i)\\
&-\nabla f_i(\mathbf{z}_{t-1}^i,\xi_{t-1}^i)-\nabla\mathbf{f}_i(\mathbf{z_t^i}))\\
&=\beta(\nabla f_i(\mathbf{z}_t^i,\xi_{t-1}^i)-\nabla\mathbf{f}_i(\mathbf{z_t^i}))\\
&+(1-\beta)(\nabla f_i(\mathbf{z}_t^i,\xi_{t-1}^i)-\nabla f_i(\mathbf{z}_{t-1}^i,\xi_{t-1}^i)\\
&+\nabla\mathbf{f}_i(\mathbf{z_{t-1}^i})-\nabla\mathbf{f}_i(\mathbf{z_t^i}))+(1-\beta)(\mathbf{v}_{t-1}^i-\nabla\mathbf{f}_i(\mathbf{z_{t-1}^i})).
\end{aligned}$$

Let $\mathbf{s}_t=\sum_i\nabla f_i(\mathbf{z}_t^i,\xi_{t-1}^i)-\nabla\mathbf{f}_i(\mathbf{z_t^i})$ and $\mathbf{s}_t'=\sum_i\nabla f_i(\mathbf{z}_t^i,\xi_{t-1}^i)-\nabla f_i(\mathbf{z}_{t-1}^i,\xi_{t-1}^i)+\nabla\mathbf{f}_i(\mathbf{z_{t-1}^i})-\nabla\mathbf{f}_i(\mathbf{z_t^i})$. Summing over i from 1 to n and taking the expectation conditioned on $\mathcal{F}_t$ gives

$$\begin{aligned}
&\mathbb{E}[\|\mathbf{v}_t-\nabla\mathbf{f}(\mathbf{z_t})\|^2|\mathcal{F}_t]\\
&=(1-\beta)^2\|\mathbf{v}_{t-1}-\nabla\mathbf{f}(\mathbf{z_{t-1}})\|^2+\mathbb{E}[\|\beta\mathbf{s}_t+(1-\beta)\mathbf{s}_t'\|^2|\mathcal{F}_t]\\
&+2\mathbb{E}[\langle(1-\beta)(\mathbf{v}_{t-1}-\nabla\mathbf{f}(\mathbf{z_{t-1}})),\beta\mathbf{s}_t+(1-\beta)\mathbf{s}_t'\rangle|\mathcal{F}_t]\\
&\overset{(3a)}{\leq}(1-\beta)^2\|\mathbf{v}_{t-1}-\nabla\mathbf{f}(\mathbf{z_{t-1}})\|^2+\mathbb{E}[\|\beta\mathbf{s}_t+(1-\beta)\mathbf{s}_t'\|^2|\mathcal{F}_t]\\
&\overset{(3b)}{\leq}(1-\beta)^2\|\mathbf{v}_{t-1}-\nabla\mathbf{f}(\mathbf{z_{t-1}})\|^2+2\beta^2\mathbb{E}[\|\mathbf{s}_t\|^2|\mathcal{F}_t]\\
&+2(1-\beta)^2\mathbb{E}[\|\mathbf{s}_t'\|^2|\mathcal{F}_t]\\
&\overset{(3c)}{=}(1-\beta)^2\|\mathbf{v}_{t-1}-\nabla\mathbf{f}(\mathbf{z_{t-1}})\|^2+2\beta^2\mathbb{E}[\|\mathbf{s}_t\|^2|\mathcal{F}_t]\\
&+2(1-\beta)^2\mathbb{E}[\|\sum_i\nabla f_i(\mathbf{z}_t^i,\xi_{t-1}^i)-\nabla f_i(\mathbf{z}_{t-1}^i,\xi_{t-1}^i)\|^2|\mathcal{F}_t],
\end{aligned}$$

where $(3a)$ is due to Assumption 1, $(3b)$ is due to the Cauchy-Schwarz inequality and $(3c)$ is due to the conditional variance decomposition. The upper bound on the unconditional expectation can then be derived as

$$\begin{aligned}
&\mathbb{E}[\|\mathbf{v}_t-\nabla\mathbf{f}(\mathbf{z_t})\|^2]\\
&\leq(1-\beta)^2\mathbb{E}[\|\mathbf{v}_{t-1}-\nabla\mathbf{f}(\mathbf{z_{t-1}})\|^2]+2\beta^2n\bar{\nu}^2\\
&+2(1-\beta)^2L^2\mathbb{E}[\|\mathbf{z}_t-\mathbf{z}_{t-1}\|^2]\\
&\overset{(3d)}{\leq}(1-\beta)^2\mathbb{E}[\|\mathbf{v}_{t-1}-\nabla\mathbf{f}(\mathbf{z_{t-1}})\|^2]+2\beta^2n\bar{\nu}^2+6(1-\beta)^2\\
&L^2(\mathbb{E}[\|\mathbf{z}_t-\bar{\mathbf{x}}_t\|^2+\|\bar{\mathbf{x}}_t-\bar{\mathbf{x}}_{t-1}\|^2+\|\mathbf{z}_{t-1}-\bar{\mathbf{x}}_{t-1}\|^2])\\
&=(1-\beta)^2\mathbb{E}[\|\mathbf{v}_{t-1}-\nabla\mathbf{f}(\mathbf{z_{t-1}})\|^2]+2\beta^2n\bar{\nu}^2\\
&+6(1-\beta)^2L^2\alpha^2\mathbb{E}[\|\bar{\mathbf{v}}_{t-1}\|^2]\\
&+6(1-\beta)^2L^2(\mathbb{E}[\|\mathbf{z}_t-\bar{\mathbf{x}}_t\|^2+\|\mathbf{z}_{t-1}-\bar{\mathbf{x}}_{t-1}\|^2])\\
&\overset{(3e)}{\leq}(1-\beta)^2\mathbb{E}[\|\mathbf{v}_{t-1}-\nabla\mathbf{f}(\mathbf{z_{t-1}})\|^2]+2\beta^2n\bar{\nu}^2\\
&+6(1-\beta)^2L^2\alpha^2\mathbb{E}[\|\bar{\mathbf{v}}_{t-1}\|^2]\\
&+12(1-\beta)^2L^2\phi_m(n+1)(\mathbb{E}[\mathfrak{L}^2(\mathbf{z}_t,\phi_t)+\mathfrak{L}^2(\mathbf{z}_{t-1},\phi_{t-1})]),
\end{aligned}\tag{33}$$

where $(3d)$ is due to Cauchy-Schwarz inequality and $(3e)$ is from (32). This completes the proof of the first inequality in Lemma 3. The upper bound on the averaged version can be derived using the same technique by replacing $\mathbb{E}[\|\mathbf{v}_t-\nabla\mathbf{f}(\mathbf{z_t})\|^2]$ with $\mathbb{E}[\|\bar{\mathbf{v}}_t-\nabla\bar{\mathbf{f}}(\mathbf{z_t})\|^2]$.

D. Proof of Lemma 4

Since the global function is L-smooth, for $\alpha\in(0,\frac{1}{2L}]$ it holds that

$$\begin{aligned}
&f(\bar{\mathbf{x}}_{t+1})\overset{(4a)}{\leq}f(\bar{\mathbf{x}}_t)+<\nabla f(\bar{\mathbf{x}}_t),\bar{\mathbf{x}}_{t+1}-\bar{\mathbf{x}}_t>+\frac{L}{2}\|\bar{\mathbf{x}}_{t+1}-\bar{\mathbf{x}}_t\|^2\\
&\overset{(4b)}{\leq}f(\bar{\mathbf{x}}_t)-\alpha<\nabla f(\bar{\mathbf{x}}_t),\bar{\mathbf{v}}_t>+\frac{L\alpha^2}{2}\|\bar{\mathbf{v}}_t\|^2\\
&\overset{(4c)}{\leq}f(\bar{\mathbf{x}}_t)-\frac{\alpha}{2}\|\nabla f(\bar{\mathbf{x}}_t)\|^2-(\frac{\alpha}{2}-\frac{L\alpha^2}{2})\|\bar{\mathbf{v}}_t\|^2\\
&+\frac{\alpha}{2}\|\bar{\mathbf{v}}_t-\nabla f(\bar{\mathbf{x}}_t)\|^2\\
&\overset{(4d)}{\leq}f(\bar{\mathbf{x}}_t)-\frac{\alpha}{2}\|\nabla f(\bar{\mathbf{x}}_t)\|^2-(\frac{\alpha}{2}-\frac{L\alpha^2}{2})\|\bar{\mathbf{v}}_t\|^2\\
&+\alpha\|\bar{\mathbf{v}}_t-\nabla\bar{\mathbf{f}}(\mathbf{z_t})\|^2+\alpha\|\nabla\bar{\mathbf{f}}(\mathbf{z_t})-\nabla f(\bar{\mathbf{x}}_t)\|^2\\
&\overset{(4e)}{\leq}f(\bar{\mathbf{x}}_t)-\frac{\alpha}{2}\|\nabla f(\bar{\mathbf{x}}_t)\|^2-\frac{\alpha}{4}\|\bar{\mathbf{v}}_t\|^2+\alpha\|\bar{\mathbf{v}}_t-\nabla\bar{\mathbf{f}}(\mathbf{z_t})\|^2\\
&+\frac{\alpha L^2}{n}\|\mathbf{z}_t-\bar{\mathbf{x}}_t\|^2,
\end{aligned}$$

where $(4a)$ is due to L-smoothness, $(4b)$ follows from the update of $\mathbf{x}_t$, $(4c)$ is due to the perfect square formula, $(4d)$ is due to the Cauchy-Schwarz inequality and $(4e)$ is due to the smoothness assumption and the range of α. Moving $\|\nabla f(\bar{\mathbf{x}}_t)\|^2$ to the left side yields

$$\begin{aligned}
\|\nabla f(\bar{\mathbf{x}}_t)\|^2 &\leq \frac{2(f(\bar{\mathbf{x}}_{t+1})-f(\bar{\mathbf{x}}_t))}{\alpha}-\frac{1}{2}\|\bar{\mathbf{v}}_t\|^2\\
&+2\|\bar{\mathbf{v}}_t-\nabla\bar{\mathbf{f}}(\mathbf{z_t})\|^2+\frac{2L^2}{n}\|\mathbf{z}_t-\bar{\mathbf{x}}_t\|^2.
\end{aligned}$$

This article has been accepted for publication in IEEE Transactions on Automatic Control. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAC.2024.3479888

By taking the telescoping sum over t from 0 to T,

$$\begin{aligned}
&\sum_{t=0}^{T-1}\|\nabla f(\bar{\mathbf{x}}_t)\|^2 \leq \frac{2(f(\bar{\mathbf{x}}_0)-f(\bar{\mathbf{x}}_T))}{\alpha}-\frac{1}{2}\sum_{t=0}^{T-1}\|\bar{\mathbf{v}}_t\|^2\\
&+2\sum_{t=0}^{T-1}\|\bar{\mathbf{v}}_t-\nabla\bar{\mathbf{f}}(\mathbf{z_t})\|^2+\frac{2L^2}{n}\sum_{t=0}^{T-1}\|\mathbf{z}_t-\bar{\mathbf{x}}_t\|^2\\
&\leq \frac{2(f(\bar{\mathbf{x}}_0)-f(\bar{\mathbf{x}}_T))}{\alpha}-\frac{1}{2}\sum_{t=0}^{T-1}\|\bar{\mathbf{v}}_t\|^2+2\sum_{t=0}^{T-1}\|\bar{\mathbf{v}}_t-\nabla\bar{\mathbf{f}}(\mathbf{z_t})\|^2\\
&+\frac{4L^2(n+1)\phi_m}{n}\sum_{t=0}^{T-1}\mathfrak{L}^2(\mathbf{z}_t,\phi_t)
\end{aligned}\tag{34}$$

where the last inequality is due to (32). The proof is completed by taking expectation of both sides of (34).

E. Proof of Theorem 1

Using the fact that $\frac{1}{1-(1-\beta)^2}\leq\frac{1}{\beta}$ for $\beta\in(0,1)$, leveraging the recursive error bounds in inequality $(3e)$ in (33), and applying telescoping leads to

$$\begin{aligned}
&\sum_{t=0}^{T}\mathbb{E}[\|\mathbf{v}_t-\nabla\mathbf{f}(\mathbf{z_t})\|^2]\leq\frac{1}{\beta}\mathbb{E}[\|\mathbf{v}_0-\nabla\mathbf{f}(\mathbf{z_0})\|^2]+2\beta n\bar{\nu}^2T\\
&+\frac{1}{\beta}6(1-\beta)^2L^2\alpha^2\sum_{t=1}^{T}\mathbb{E}[\|\bar{\mathbf{v}}_{t-1}\|^2]\\
&+\frac{1}{\beta}24(1-\beta)^2L^2\phi_m(n+1)\sum_{t=0}^{T}\mathbb{E}[\mathfrak{L}^2(\mathbf{z}_t,\phi_0)].
\end{aligned}$$

Moreover, taking the telescoping sum over Lemma 1 leads to

$$\begin{aligned}
&\sum_{t=0}^{T}\mathbb{E}\mathfrak{L}^2(\mathbf{z}_t,\phi_t)\leq\frac{1}{1-(1+\delta^2)/2}\mathfrak{L}^2(\mathbf{z}_0,\phi_0)\\
&+\frac{1}{1-(1+\delta^2)/2}\frac{2\delta^2\alpha^2}{1-\delta^2}\sum_{t=0}^{T}\mathbb{E}\mathfrak{L}^2(\mathbf{h}_t,\phi_0)\\
&=\frac{2}{1-\delta^2}\mathfrak{L}^2(\mathbf{z}_0,\phi_0)+\frac{4\delta^2\alpha^2}{(1-\delta^2)^2}\sum_{t=0}^{T}\mathbb{E}\mathfrak{L}^2(\mathbf{h}_t,\phi_0),
\end{aligned}$$

while taking the telescoping sum over Lemma 2 leads to

$$\begin{aligned}
&\sum_{t=0}^{T}\mathbb{E}[\mathfrak{L}^2(\mathbf{h}_t,\phi_t)]\\
&\leq\frac{2}{1-\delta^2}\mathbb{E}[\|\check{\mathbf{h}}_0\|^2]+\frac{4}{1-\delta^2}\frac{72\delta^2\alpha^2L^2}{(1-\delta^2)^2}\|Y^{-1}\|^2\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{\mathbf{v}}_t\|^2]\\
&+\frac{4\delta^2}{(1-\delta^2)^2}\|Y^{-1}\|^2144L^2\phi_m(n+1)\sum_{t=0}^{T}\mathbb{E}\mathfrak{L}^2(\mathbf{z}_t,\phi_t)\\
&+\frac{96\delta^2}{(1-\delta^2)^2}\|Y^{-1}\|^2L^2\beta^2\sum_{t=0}^{T}\mathbb{E}[\|\mathbf{v}_t-\nabla\mathbf{f}(\mathbf{z_t})\|^2]\\
&+\frac{4\delta^2}{(1-\delta^2)^2}\|Y^{-1}\|^224\beta^2n\bar{\nu}^2T.
\end{aligned}$$

Utilizing the upper bounds above, we can derive the range of the step size α such that the upper bound on $\sum_{t=0}^{T}\mathbb{E}[\mathfrak{L}^2(\mathbf{h}_t,\phi_t)]$ is independent of the other error terms. Letting $\frac{2304\delta^4\alpha^2}{(1-\delta^2)^4}\|Y^{-1}\|^2L^2\phi_m(n+1)+\frac{9216\delta^4\alpha^2}{(1-\delta^2)^4}\|Y^{-1}\|^2L^4\phi_m(n+1)<\frac{1}{2}$, we have

$$\begin{aligned}
&\sum_{t=0}^{T}\mathbb{E}[\mathfrak{L}^2(\mathbf{h}_t,\phi_t)]\\
&\leq\frac{4}{1-\delta^2}\mathbb{E}[\mathfrak{L}^2(\mathbf{h}_0,\phi_0)]+\frac{576\delta^2\alpha^2L^2}{(1-\delta^2)^3}\|Y^{-1}\|^2\sum_{t=0}^{T-1}\mathbb{E}[\|\bar{\mathbf{v}}_t\|^2]\\
&+\frac{2304\delta^2}{(1-\delta^2)^3}\|Y^{-1}\|^2L^2(n^2+1)\mathfrak{L}^2(\mathbf{z}_0,\phi_0)\\
&+\frac{192\delta^2}{(1-\delta^2)^2}\|Y^{-1}\|^2\beta^2n\bar{\nu}^2T\\
&+\frac{192\delta^2}{(1-\delta^2)^2}\|Y^{-1}\|^2L^2\beta^2[\frac{1}{\beta}\mathbb{E}[\|\mathbf{v}_0-\nabla\mathbf{f}(\mathbf{z_0})\|^2]+2\beta n\bar{\nu}^2T\\
&+\frac{1}{\beta}6(1-\beta)^2L^2\alpha^2\sum_{t=1}^{T}\mathbb{E}[\|\bar{\mathbf{v}}_{t-1}\|^2]\\
&+\frac{1}{\beta}24(1-\beta)^2L^2\phi_m(n+1)\frac{2}{1-\delta^2}\mathfrak{L}^2(\mathbf{z}_0,\phi_0)].
\end{aligned}$$

Initializing by all-zero models and identifying the range of step size α such that the coefficient for the term $\sum_{t=0}^{T-1}\mathbb{E}\|\bar{\mathbf{v}}_t\|^2$ is negative completes the proof of Theorem 1.

F. Proof of Corollary 1.1

To prove the corollary, we note that

$$\begin{aligned}
\mathbb{E}[\|\mathbf{v}_0-\nabla\mathbf{f}(\mathbf{z_0})\|^2&=\sum_{i=1}^{n}\mathbb{E}[\|\frac{1}{b}\sum_{r=1}^{b}(\mathbf{g}_i(\mathbf{z}^{\mathbf{i}}_{\mathbf{0}},\xi^i_{0,r})-\nabla f_i(\mathbf{z}^{\mathbf{i}}_{\mathbf{0}}))\|^2]\\
&=\frac{1}{b^2}\sum_{i=1}^{n}\sum_{r=1}^{b}\mathbb{E}[\|(\mathbf{g}_i(\mathbf{z}^{\mathbf{i}}_{\mathbf{0}},\xi^i_{0,r})-\nabla f_i(\mathbf{z}^{\mathbf{i}}_{\mathbf{0}}))\|^2]\\
&\leq\frac{n\bar{\nu}^2}{b}
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}[\|\bar{\mathbf{v}}_0-\nabla\bar{\mathbf{f}}(\mathbf{z_0})\|^2&=\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^{n}\frac{1}{b}\sum_{r=1}^{b}(\mathbf{g}_i(\mathbf{z}^{\mathbf{i}}_{\mathbf{0}},\xi^i_{0,r})-\nabla f_i(\mathbf{z}^{\mathbf{i}}_{\mathbf{0}}))\|^2]\\
&\leq\frac{\bar{\nu}^2}{nb}.
\end{aligned}$$

From the result of Theorem 1,

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\bar{\mathbf{x}}_t)\|^2\leq\frac{2\mathbb{E}(f(\bar{\mathbf{x}}_0)-f(\bar{\mathbf{x}}_T))}{\alpha T}\\
&+\frac{2}{\beta T}\mathbb{E}[\|\bar{\mathbf{v}}_0-\nabla\bar{\mathbf{f}}(\mathbf{z_0})\|^2]+4\beta\bar{\nu}^2+\mathcal{O}(\frac{\alpha^2}{\beta}(\frac{1}{T}+\beta^2))\\
&\leq\frac{2\mathbb{E}(f(\bar{\mathbf{x}}_0)-f(\bar{\mathbf{x}}_T))}{\alpha T}+\frac{2}{\beta T}\frac{\bar{\nu}^2}{nb}+4\beta\bar{\nu}^2+\mathcal{O}(\frac{\alpha^2}{\beta}(\frac{1}{T}+\beta^2)).
\end{aligned}$$

By letting $\alpha=\mathcal{O}(\frac{1}{n^{1/2}T^{1/3}})$, $\beta=\mathcal{O}(\frac{1}{T^{2/3}})$, and $b=\mathcal{O}(\frac{T^{1/3}}{n})$, we conclude that

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(\bar{\mathbf{x}}_t)\|^2\leq\mathcal{O}(\frac{1}{T^{2/3}}),\tag{35}$$

which completes the proof of the corollary.

G. Proof of Lemma 5

Revisiting the proof of Lemma 4 (specifically, expressions $(4a)$ – $(4e)$) to introduce the PL condition yields

$$\begin{aligned}f(\bar{\mathbf{x}}_{t+1}) &\leq f(\bar{\mathbf{x}}_t) - < \nabla f(\bar{\mathbf{x}}_t), \bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_t > + \frac{L}{2}\|\bar{\mathbf{x}}_{t+1} - \bar{\mathbf{x}}_t\|^2 \\ &\leq f(\bar{\mathbf{x}}_t) - \frac{\alpha}{2}\|\nabla f(\bar{\mathbf{x}}_t)\|^2 - \frac{\alpha}{4}\|\bar{\mathbf{v}}_t\|^2 + \alpha\|\bar{\mathbf{v}}_t - \nabla\bar{\mathbf{f}}(\mathbf{z_t})\|^2 \\ &+ \frac{\alpha L^2}{n}\|\mathbf{z}_t - \bar{\mathbf{x}}_t\|^2 \\ &\overset{(5a)}{\leq} f(\bar{\mathbf{x}}_t) - \frac{\alpha\mu}{2}(f(\bar{\mathbf{x}}_t) - f^*) - \frac{\alpha}{4}\|\bar{\mathbf{v}}_t\|^2 \\ &+ \alpha\|\bar{\mathbf{v}}_t - \nabla\bar{\mathbf{f}}(\mathbf{z_t})\|^2 + \frac{\alpha L^2}{n}\|\mathbf{z}_t - \bar{\mathbf{x}}_t\|^2,\end{aligned}$$

where $(5a)$ is due to the PL condition. Subtracting f^* from both sides and taking the expectation results in

$$\begin{aligned}\mathbb{E}[f(\bar{\mathbf{x}}_{t+1}) - f^*] &\leq \mathbb{E}[(1 - \frac{\alpha\mu}{2})(f(\bar{\mathbf{x}}_t) - f^*) - \frac{\alpha}{4}\|\bar{\mathbf{v}}_t\|^2 \\ &\quad + \alpha\|\bar{\mathbf{v}}_t - \nabla\bar{\mathbf{f}}(\mathbf{z_t})\|^2 + \frac{\alpha L^2}{n}\|\mathbf{z}_t - \bar{\mathbf{x}}_t\|^2] \\ &\leq \mathbb{E}[(1 - \frac{\alpha\mu}{2})(f(\bar{\mathbf{x}}_t) - f^*) - \frac{\alpha}{4}\|\bar{\mathbf{v}}_t\|^2 \\ &\quad + \frac{\alpha}{n}\|\mathbf{v}_t - \nabla\mathbf{f}(\mathbf{z_t})\|^2 + \frac{\alpha L^2\phi_m(2n+2)}{n}\mathfrak{L}^2(\mathbf{z}_t, \phi_t)].\end{aligned}$$

Having accumulated recursive error bounds for each of the four errors impacting the linear inequality system at iteration k, we obtain

$$\mathbf{u}_{k+1} \leq C\mathbf{u}_k + \mathbf{r}_k.$$

H. Proof of Lemma 6

Recall that, as we argued in the discussion following Lemma 5, the existence of $\mathbf{x} > 0$ such that $C\mathbf{x} < \mathbf{x}$ implies $\rho(C) < 1$. Building on a similar idea, here we aim to find a range for the step size α and a specific vector $\mathbf{x} > 0$ such that $C\mathbf{x} \leq (1 - \frac{\alpha\mu}{4})\mathbf{x}$, which subsequently implies $\rho(C) \leq 1 - \frac{\alpha\mu}{4}$. To this end, we construct the following inequalities by re-writing $C\mathbf{x} \leq (1 - \frac{\alpha\mu}{4})\mathbf{x}$ in scalar format for 4-dimensional $\mathbf{x}$ (recall the definition of C in (25)):

$$\begin{aligned}&\frac{1+\delta^2}{2}x_1 + \frac{2\delta^2\alpha^2}{1-\delta^2}x_2 \leq (1 - \frac{\alpha\mu}{4})x_1, \\ &\frac{144\delta^2\|Y^{-1}\|^2L^2\phi_m(n+1)}{1-\delta^2}x_1 + (\frac{288\delta^4\|Y^{-1}\|^2L^2\phi_m\alpha^2(n+1)}{1-\delta^2} \\ &+ \frac{1+\delta^2}{2})x_2 + \frac{24\delta^2\beta^2\|Y^{-1}\|^2}{1-\delta^2}x_3 \leq (1 - \frac{\alpha\mu}{4})x_2, \\ &24(1-\beta)^2L^2\phi_m(n+1)x_1 + 24(1-\beta)^2L^2\phi_m(n+1)\frac{\delta^2\alpha^2}{1-\delta^2}x_2 \\ &+ (1-\beta)^2x_3 \leq (1 - \frac{\alpha\mu}{4})x_3, \\ &\frac{2\alpha L^2\phi_m(n+1)}{n}x_1 + \frac{\alpha}{n}x_3 + (1 - \frac{\alpha\mu}{2})x_4 \overset{(6a)}{\leq} (1 - \frac{\alpha\mu}{4})x_4.\end{aligned}$$

By rearranging the first three inequalities above, we obtain the following inequalities on x_1/x_2 and x_3/x_2:

$$\frac{\frac{2\delta^2\alpha^2}{1-\delta^2}}{\frac{1-\delta^2}{2} - \frac{\alpha\mu}{4}} \leq \frac{x_1}{x_2}, \tag{36}$$

$$\frac{x_1}{x_2} \leq \frac{\frac{1}{2}(\frac{1-\delta^2}{2} - \frac{\alpha\mu}{4} - \frac{288\delta^4\|Y^{-1}\|^2L^2\phi_m\alpha^2(n+1)}{1-\delta^2})}{\frac{144\delta^2\|Y^{-1}\|^2L^2\phi_m(n+1)}{1-\delta^2}}, \tag{37}$$

$$\frac{24(1-\beta)^2L^2\phi_m(n+1)\frac{2\delta^2\alpha^2}{1-\delta^2}}{\frac{1}{2}(1-(1-\beta)^2 - \frac{\alpha\mu}{4})(\frac{1-\delta^2}{2} - \frac{\alpha\mu}{4})} \leq \frac{x_3}{x_2}, \tag{38}$$

$$\frac{x_3}{x_2} \leq \frac{\frac{1}{2}(\frac{1-\delta^2}{2} - \frac{\alpha\mu}{4} - \frac{288\delta^4\|Y^{-1}\|^2L^2\phi_m\alpha^2(n+1)}{1-\delta^2})}{\frac{24\delta^2\beta^2\|Y^{-1}\|^2}{1-\delta^2}}. \tag{39}$$

Finally, we set $x_2 = 1$ and solve recursively for x_1, x_3 and x_4 via $(6a)$ and (36)-(39). The desired inequality $C\mathbf{x} \leq (1 - \frac{\alpha\mu}{4})\mathbf{x}$ holds for the following choice of $\mathbf{x}$:

$$\begin{aligned}x_1 &= \frac{\frac{2\delta^2\alpha^2}{1-\delta^2}}{\frac{1-\delta^2}{2} - \frac{\alpha\mu}{4}}, \\ x_2 &= 1, \\ x_3 &= \frac{96(1-\beta)^2L^2(n^2+1)\frac{\delta^2\alpha^2}{1-\delta^2}}{(1-(1-\beta)^2 - \frac{\alpha\mu}{4})(\frac{1-\delta^2}{2} - \frac{\alpha\mu}{4})}, \\ x_4 &= \frac{4}{\alpha\mu}[\frac{2\alpha L^2\phi_m(n+1)}{n}x_1 + \frac{\alpha}{n}x_3].\end{aligned}$$

Therefore, given the above construction, there exist positive x_1, x_2, x_3 and x_4 such that $\rho(C) \leq 1 - \frac{\alpha\mu}{4}$. The corresponding range of α can then be determined via (36)-(39).

I. Proof of Lemma 7

Recall the definition of C in (25). To facilitate upcoming analysis of the linear system inequality, let us start by computing the inverse of $I_4 - C$, i.e.,

$$(I_4 - C)^{-1} = \overline{I_4 - C}/\det(I_4 - C),$$

where $\overline{I_4 - C}$ is the adjugate matrix of $I_4 - C$. To compute this inverse, we first find and bound the determinant of $I_4 - C$,

$$\begin{aligned}&\det(I_4 - C) \\ &= \frac{\alpha\mu}{2}\{(\frac{(1-\delta^2)^2}{4} - \frac{144\delta^4\alpha^2L^2\|Y^{-1}\|^2\phi_m(n+1)}{(1-\delta^2)})(2\beta - \beta^2) \\ &- 24(1-\beta)^2L^2\phi_m(n+1)\frac{48\delta^4\alpha^2\beta^2\|Y^{-1}\|^2}{(1-\delta^2)^2} \\ &- \frac{576\delta^4\alpha^2\|Y^{-1}\|^2L^2\phi_m(n+1)(1-(1-\beta)^2)}{(1-\delta^2)^2} \\ &- 288(1-\beta)^2L^2\phi_m(n+1)\delta^4\alpha^2\beta^2\frac{\|Y^{-1}\|^2}{1-\delta^2}\} \\ &\overset{(7a)}{\geq} \frac{\alpha\mu}{6}(\frac{(1-\delta^2)^2}{4} - \frac{144\delta^4\alpha^2L^2\|Y^{-1}\|^2\phi_m(n+1)}{(1-\delta^2)})(2\beta - \beta^2).\end{aligned}$$

The lower bound $(7a)$ is achieved if the learning rate/stepsize parameters α and β are chosen such that

$$\begin{aligned}&24(1-\beta)^2L^2\phi_m(n+1)\frac{48\delta^4\alpha^2\beta^2\|Y^{-1}\|^2}{(1-\delta^2)^2} \\ &+ \frac{576\delta^4\alpha^2\|Y^{-1}\|^2L^2\phi_m(n+1)(2\beta - \beta^2)}{(1-\delta^2)^2} \\ &+ 288(1-\beta)^2L^2\phi_m(n+1)\delta^4\alpha^2\beta^2\frac{\|Y^{-1}\|^2}{1-\delta^2} \\ &\leq (\frac{(1-\delta^2)^2}{6} - \frac{96\delta^4\alpha^2L^2\|Y^{-1}\|^2\phi_m(n+1)}{1-\delta^2})(2\beta - \beta^2).\end{aligned}$$

This also implies that

$$\frac{288\delta^4\alpha^2L^2\|Y^{-1}\|^2\phi_m(n+1)}{(1-\delta^2)} \leq \frac{(1-\delta^2)^2}{8}.$$

By invoking $(7a)$, we can show that $\det(I_4 - C) \geq \frac{\alpha\mu\beta(1-\delta^2)^2}{96}$. Next, we note that the adjugate matrix needed to specify the inverse of $I_4 - C$ has entries that satisfy the following (in)equalities:

$$\begin{aligned}
[\overline{I_4 - C}]_{1,2} &= \frac{2\alpha^3\beta\mu\delta^2}{1-\delta^2} - \frac{\alpha^3\beta^2\mu\delta^2}{1-\delta^2} \\
[\overline{I_4 - C}]_{1,3} &= \frac{24\|Y^{-1}\|^2\alpha^3\beta^2\mu\delta^4}{(1-\delta^2)^2} \\
[\overline{I_4 - C}]_{2,2} &= \frac{\alpha\mu\beta(1-\delta^2)}{2} - \frac{\alpha\mu\beta^2(1-\delta^2)}{4} \\
[\overline{I_4 - C}]_{2,3} &= 6\|Y^{-1}\|^2\alpha\beta^2\mu\delta^2 \\
[\overline{I_4 - C}]_{3,2} &\leq \frac{30L^2\alpha^3\mu\delta^2\phi_m(1+\beta^2+n+\beta^2 n)}{1-\delta^2} \\
[\overline{I_4 - C}]_{3,3} &\leq \frac{\alpha\mu}{8} \\
[\overline{I_4 - C}]_{4,2} &\leq \frac{L^2\alpha^3\delta^2\phi_m(60+56\beta^2+60n+56\beta^2 n)}{n(1-\delta^2)} \\
[\overline{I_4 - C}]_{4,3} &\leq \frac{\alpha(1-\delta^2)^2}{4n} \\
&\quad + \frac{L^2\|Y^{-1}\|^2\alpha^3\delta^4\phi_m(96\beta^2+144\delta^2+96\beta^2 n)}{n(1-\delta^2)^2}.
\end{aligned}$$

Recall (27) which states that for all $k \in [1, T]$,

$$\begin{aligned}
\mathbf{u}_k \leq C^k\mathbf{u}_0 + \sum_{t=0}^{k-1} C^t\mathbf{r}_k &\leq C^k\mathbf{u}_0 + \sum_{t=0}^{k-1} C^t\mathbf{r}_{k_m} \\
&\leq C^k\mathbf{u}_0 + (I_4 - C)^{-1}\mathbf{r}_{k_m},
\end{aligned}$$

where $\mathbf{r}_{k_m} = \max_{k\in[1,T]}\mathbf{r}_k$ and $k_m = \mathrm{argmax}_{k\in[1,T]}\mathbf{r}_k$.

Since $\mathbf{r}_k$ is a function of $E[\|\bar{\mathbf{v}}_k\|^2]$, we proceed by deriving an upper bound on $E[\|\bar{\mathbf{v}}_k\|^2]$. Following the proof of Lemma 5,

$$\begin{aligned}
&\frac{\alpha}{4}\mathbb{E}[\|\bar{\mathbf{v}}_k\|^2] \leq \mathbb{E}[(1-\frac{\alpha\mu}{2})(f(\bar{\mathbf{x}}_k) - f^*) \\
&+ \frac{\alpha}{n}\|\mathbf{v}_k - \nabla\mathbf{f}(\mathbf{z_k})\|^2 + \frac{\alpha L^2\phi_m(2n+2)}{n}\mathfrak{L}^2(\mathbf{z}_t, \phi_t)] \\
&\overset{(7b)}{\leq} C_{exp} + (1-\frac{\alpha\mu}{2})[(I_4 - C)^{-1}\mathbf{r}_k]_4 \\
&+ \frac{\alpha}{n}[(I_4 - C)^{-1}\mathbf{r}_k]_3 + \frac{\alpha L^2\phi_m(2n+2)}{n}[(I_4 - C)^{-1}\mathbf{r}_k]_1,
\end{aligned}$$

where

$$\begin{aligned}
C_{exp} = \max_{k_m}\{&(1-\frac{\alpha\mu}{2})[C^{k_m}\mathbf{u}_0]_4 + \frac{\alpha}{n}[C^{k_m}\mathbf{u}_0]_3 \\
&+ \frac{\alpha L^2\phi_m(2n+2)}{n}[C^{k_m}\mathbf{u}_0]_1\}.
\end{aligned}$$

We leverage the error bounds in Lemmas 1, 2, 3 and 5 and the computation of $(I_4 - C)^{-1}$ in $(7b)$ to establish

$$\begin{aligned}
&\frac{1}{4}\mathbb{E}[\|\bar{\mathbf{v}}_k\|^2] \\
&\leq \frac{C_{exp}}{\alpha} + \frac{96(1-\frac{\alpha\mu}{2})}{\mu\beta(1-\delta^2)^2}[\frac{L^2\alpha\delta^2\phi_m(60+56\beta^2+60n+56\beta^2 n)}{n(1-\delta^2)} \\
&\frac{8\delta^2}{1-\delta^2}\|Y^{-1}\|^2(9\alpha^2L^2\mathbb{E}[\|\bar{\mathbf{v}}_k\|^2] + 3\beta^2\bar{\nu}^2 n) \\
&+ (\frac{(1-\delta^2)^2}{4\alpha n} + \frac{L^2\|Y^{-1}\|^2\alpha\delta^4\phi_m(96\beta^2+144\delta^2+96\beta^2 n)}{n(1-\delta^2)^2}) \\
&(6(1-\beta)^2L^2\alpha^2\mathbb{E}[\|\bar{\mathbf{v}}_k\|^2] + 2\beta^2\bar{\nu}^2 n)] \\
&+ \frac{1}{n}\frac{96}{\beta(1-\delta^2)^2} \times [\frac{30L^2\alpha^2\delta^2\phi_m(1+\beta^2+n+\beta^2 n)}{1-\delta^2}\frac{8\delta^2}{1-\delta^2} \\
&\|Y^{-1}\|^2(9\alpha^2L^2\mathbb{E}[\|\bar{\mathbf{v}}_k\|^2] + 3\beta^2\bar{\nu}^2 n) \\
&+ \frac{1}{8}(6(1-\beta)^2L^2\alpha^2\mathbb{E}[\|\bar{\mathbf{v}}_k\|^2] + 2\beta^2\bar{\nu}^2 n)] \\
&+ \frac{L^2\phi_m(2n+2)}{n}\frac{96}{(1-\delta^2)^2} \times [(\frac{2\alpha^2\delta^2}{1-\delta^2} - \frac{\alpha^2\beta\delta^2}{1-\delta^2})\frac{8\delta^2}{1-\delta^2}\|Y^{-1}\|^2 \\
&(9\alpha^2L^2\mathbb{E}[\|\bar{\mathbf{v}}_k\|^2] + 3\beta^2\bar{\nu}^2 n) \\
&+ \frac{24\|Y^{-1}\|^2\alpha^2\beta\delta^4}{(1-\delta^2)^2}(6(1-\beta)^2L^2\alpha^2\mathbb{E}[\|\bar{\mathbf{v}}_k\|^2] + 2\beta^2\bar{\nu}^2 n)].
\end{aligned}$$

Collecting the terms with $\mathbb{E}[\|\bar{\mathbf{v}}_k\|^2]$ on the right-hand side and imposing

$$\begin{aligned}
&(1-\frac{\alpha\mu}{2})\frac{96}{\mu\beta(1-\delta^2)^2}\times \\
&[\frac{9L^4\alpha^3\delta^2\phi_m(60+56\beta^2+60n+56\beta^2 n)}{(1-\delta^2)}\frac{8\delta^2}{1-\delta^2}\|Y^{-1}\|^2 \\
&+ (\frac{(1-\delta^2)^2}{4\alpha n} + \frac{L^2\|Y^{-1}\|^2\alpha\delta^4\phi_m(96\beta^2+144\delta^2+96\beta^2 n)}{n(1-\delta^2)^2}) \\
&(6(1-\beta)^2L^2\alpha^2)] \\
&+ \frac{96}{n\beta(1-\delta^2)^2}[\frac{2160L^4\alpha^4\delta^4\phi_m(1+\beta^2+n+\beta^2 n)}{(1-\delta^2)^2}\|Y^{-1}\|^2 \\
&+ \frac{3}{4}(1-\beta)^2L^2\alpha^2] \\
&+ \frac{96L^2\phi_m(2n+2)}{n(1-\delta^2)^2}[(\frac{8\alpha^2\delta^2}{1-\delta^2} - \frac{\alpha^2\beta\delta^2}{1-\delta^2})\frac{2\delta^2}{1-\delta^2}\|Y^{-1}\|^2 \times 9\alpha^2L^2 \\
&+ \frac{144\|Y^{-1}\|^2\alpha^2\beta\delta^4}{(1-\delta^2)^2}(1-\beta)^2L^2\alpha^2] \leq \frac{1}{8}
\end{aligned}$$

leads to

$$\begin{aligned}
&\frac{1}{8}\mathbb{E}[\|\bar{\mathbf{v}}_k\|^2] \\
&\leq \frac{C_{exp}}{\alpha} + \frac{96(1-\frac{\alpha\mu}{2})}{\mu(1-\delta^2)^2}[\frac{8L^2\alpha\delta^4\phi_m(60+56\beta^2+60n+56\beta n)}{n(1-\delta^2)^2} \\
&\|Y^{-1}\|^2 3\beta\bar{\nu}^2 n + (\frac{(1-\delta^2)^2}{4\alpha} \\
&+ \frac{L^2\|Y^{-1}\|^2\alpha\delta^4\phi_m(96\beta^2+144\delta^2+96\beta^2 n)}{(1-\delta^2)^2})2\beta^2\bar{\nu}^2] \\
&+ \frac{1}{n}\frac{96}{(1-\delta^2)^2} \times [\frac{30L^2\alpha^2\delta^2\phi_m(1+\beta^2+n+\beta^2 n)}{1-\delta^2}\frac{8\delta^2}{1-\delta^2} \\
&\|Y^{-1}\|^2 3\beta\bar{\nu}^2 n + \frac{2}{8}\beta\bar{\nu}^2 n] \\
&+ \frac{L^2\phi_m(2n+2)}{n}\frac{96}{(1-\delta^2)^2} \times [(\frac{2\alpha^2\delta^2}{1-\delta^2} - \frac{\alpha^2\beta\delta^2}{1-\delta^2})\frac{8\delta^2}{1-\delta^2} \\
&\|Y^{-1}\|^2 3\beta^2\bar{\nu}^2 n + \frac{48\|Y^{-1}\|^2\alpha^2\beta\delta^4}{(1-\delta^2)^2}\beta^2\bar{\nu}^2 n].
\end{aligned} \tag{40}$$

Since the right hand side is independent of k and T, this expression provides an upper bound on $E[\|\bar{\mathbf{v}}_k\|^2]$ w.r.t. $\bar{\nu}$ for all k, T.

J. Proof of Theorem 2

We recall conditions on the step sizes α and β used in Lemmas 1-3 and 5-7, and let the step sizes be such that

$$\alpha \leq \min\{\frac{1}{2L}, \frac{2\beta}{\mu}, \frac{1-\delta^2}{\mu}\},$$

$$576L^2\phi_m(n+1)\frac{\delta^4\alpha^2\|Y^{-1}\|^2}{(1-\delta^2)^2} \leq \frac{(1-\delta^2)\alpha\mu}{32}\times$$

$$\times(\frac{1-\delta^2}{4} - \frac{288\delta^4\|Y^{-1}\|^2L^2\alpha^2\phi_m(n+1)}{1-\delta^2}),$$

$$1728L^2\phi_m(n+1)\frac{\delta^4\alpha^2\|Y^{-1}\|^2}{(1-\delta^2)^2} + 384L^2\phi_m(n+1)\delta^4\alpha^2\frac{\|Y^{-1}\|^2}{1-\delta^2}$$

$$\leq \frac{(1-\delta^2)^2}{6},$$

and

$$\frac{96}{\mu}[\frac{3L^4\alpha\delta^4\phi_m(356+212n)\|Y^{-1}\|^2}{(1-\delta^2)^4} + \frac{3L^2\alpha}{2}]$$

$$+96[\frac{2736L^4\alpha^2\delta^4\phi_m(n+1)\|Y^{-1}\|^2}{(1-\delta^2)^4} + \frac{3L^2\alpha^2}{4(1-\delta^2)^2}] \leq \frac{\beta}{8}.$$

Then, it holds that

$$\lim_{k\to\infty}\sup\mathbb{E}[f(\bar{\mathbf{x}}_{k+1}) - f^*]$$

$$\leq \frac{96}{\mu\beta(1-\delta^2)^2}[\frac{8L^2\alpha^2\delta^4\|Y^{-1}\|^2\phi_m(60+56\beta^2+60n+56\beta^2n)}{(1-\delta^2)^2}$$

$$(9\alpha^2L^2\mathbb{E}[\|\bar{\mathbf{v}}_k\|^2] + 3\beta^2\bar{\nu}^2)$$

$$+(\frac{(1-\delta^2)^2}{4} + \frac{L^2\|Y^{-1}\|^2\alpha^2\delta^4\phi_m(96\beta^2+144\delta^2+96\beta^2n)}{(1-\delta^2)^2})$$

$$(6(1-\beta)^2L^2\alpha^2\mathbb{E}[\|\bar{\mathbf{v}}_k\|^2] + 2\beta^2\bar{\nu}^2)]$$

$$\leq (\frac{\alpha^2L^2}{4} + \frac{3L^2\alpha^2}{2\beta})\mathbb{E}[\|\bar{\mathbf{v}}_k\|^2] + \frac{96}{\mu\beta(1-\delta^2)^2}$$

$$[\frac{8L^2\alpha^2\delta^4\phi_m(60+56\beta^2+60n+56\beta^2n)}{(1-\delta^2)^2}\|Y^{-1}\|^2(3\beta^2\bar{\nu}^2)+$$

$$\frac{(1-\delta^2)^2}{4}(2\beta^2\bar{\nu}^2)+$$

$$\frac{L^2\|Y^{-1}\|^2\alpha^2\delta^4\phi_m(96\beta^2+144\delta^2+96\beta^2n)}{(1-\delta^2)^2}(2\beta^2\bar{\nu}^2)].$$

Substituting the upper bound derived for $\mathbb{E}[\|\bar{\mathbf{v}}_k\|^2]$ in (40), one can readily show that $\mathbb{E}[f(\bar{\mathbf{x}}_{k+1}) - f^*]$ decays linearly to the steady-state error given in the statement of Theorem 2.

References

[1] Yiyue Chen, Abolfazl Hashemi, and Haris Vikalo. “Accelerated Distributed Stochastic Non-Convex Optimization over Time-Varying Directed Networks”. In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2023, pp. 1–5.

[2] David Kempe, Alin Dobra, and Johannes Gehrke. “Gossip-based computation of aggregate information”. In: *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.* IEEE. 2003, pp. 482–491.

[3] Wei Ren and Randal W Beard. “Consensus seeking in multiagent systems under dynamically changing interaction topologies”. In: *IEEE Transactions on automatic control* 50.5 (2005), pp. 655–661.

[4] Angelia Nedic and Asuman Ozdaglar. “Distributed subgradient methods for multi-agent optimization”. In: *IEEE Transactions on Automatic Control* 54.1 (2009), pp. 48–61.

[5] Mahmoud Assran et al. “Stochastic gradient push for distributed deep learning”. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 344–353.

[6] Anish Acharya et al. “On the Benefits of Multiple Gossip Steps in Communication-Constrained Decentralized Federated Learning”. In: *IEEE Transactions on Parallel and Distributed Systems* (2021).

[7] Ran Xin, Soummya Kar, and Usman A Khan. “Decentralized stochastic optimization and machine learning: A unified variance-reduction framework for robust performance and fast convergence”. In: *IEEE Signal Processing Magazine* 37.3 (2020), pp. 102–113.

[8] Ran Xin, Usman A Khan, and Soummya Kar. “A hybrid variance-reduced method for decentralized stochastic non-convex optimization”. In: *arXiv preprint arXiv:2102.06752* (2021).

[9] Shi Pu and Angelia Nedić. “A distributed stochastic gradient tracking method”. In: *2018 IEEE Conference on Decision and Control (CDC)*. IEEE. 2018, pp. 963–968.

[10] Angelia Nedic, Alex Olshevsky, and Wei Shi. “Achieving geometric convergence for distributed optimization over time-varying graphs”. In: *SIAM Journal on Optimization* 27.4 (2017), pp. 2597–2633.

[11] Yiyue Chen, Abolfazl Hashemi, and Haris Vikalo. “Communication-Efficient Variance-Reduced Decentralized Stochastic Optimization over Time-Varying Directed Graphs”. In: *IEEE Transactions on Automatic Control* (2021).

[12] Muhammad I Qureshi et al. “Push-SAGA: A decentralized stochastic algorithm with variance reduction over directed graphs”. In: *IEEE Control Systems Letters* (2021).

[13] Muhammad I Qureshi et al. “S-ADDOPT: Decentralized stochastic first-order optimization over directed graphs”. In: *IEEE Control Systems Letters* 5.3 (2020), pp. 953–958.

[14] Soummya Kar, José MF Moura, and Kavita Ramanan. “Distributed parameter estimation in sensor networks: Nonlinear observation models and imperfect communication”. In: *IEEE Transactions on Information Theory* 58.6 (2012), pp. 3575–3605.

[15] Xiangru Lian et al. “Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent”. In: *arXiv preprint arXiv:1705.09056* (2017).

[16] Jinming Xu et al. “Augmented distributed gradient methods for multi-agent optimization under uncoordinated constant stepsizes”. In: *2015 54th IEEE Conference on Decision and Control (CDC)*. IEEE. 2015, pp. 2055–2060.

[17] Rudrajit Das et al. “Faster non-convex federated learning via global and local momentum”. In: *arXiv preprint arXiv:2012.04061* (2020).

[18] Ashok Cutkosky and Francesco Orabona. “Momentum-Based Variance Reduction in Non-Convex SGD”. In: *Advances in neural information processing systems* 32 (2019).

[19] Nhan H Pham et al. “ProxSARAH: An Efficient Algorithmic Framework for Stochastic Composite Nonconvex Optimization.” In: *J. Mach. Learn. Res.* 21 (2020), pp. 110–1.

[20] Cong Fang et al. “Spider: Near-optimal non-convex optimization via stochastic path integrated differential estimator”. In: *arXiv preprint arXiv:1807.01695* (2018).

[21] Angelia Nedić and Alex Olshevsky. "Distributed optimization over time-varying directed graphs". In: *IEEE Transactions on Automatic Control* 60.3 (2014), pp. 601–615.

[22] Chenguang Xi, Qiong Wu, and Usman A Khan. "On the distributed optimization over directed networks". In: *Neurocomputing* 267 (2017), pp. 508–515.

[23] Yiyue Chen, Abolfazl Hashemi, and Haris Vikalo. "Communication-efficient algorithms for decentralized optimization over directed graphs". In: *arXiv e-prints* (2020), arXiv–2005.

[24] Fakhteh Saadatniaki, Ran Xin, and Usman A Khan. "Optimization over time-varying directed graphs with row and column-stochastic matrices". In: *arXiv preprint arXiv:1810.07393* (2018).

[25] Angelia Nedich, Duong Thuy Anh Nguyen, and Duong Tung Nguyen. "AB/Push-Pull Method for Distributed Optimization in Time-Varying Directed Networks". In: *arXiv preprint arXiv:2209.06974* (2022).

[26] Duong Thuy Anh Nguyen, Duong Tung Nguyen, and Angelia Nedich. "Accelerated AB/Push-Pull Methods for Distributed Optimization over Time-Varying Directed Networks". In: *arXiv preprint arXiv:2302.01214* (2023).

[27] Léon Bottou. "Large-scale machine learning with stochastic gradient descent". In: *Proceedings of COMPSTAT'2010*. Springer, 2010, pp. 177–186.

[28] Yossi Arjevani et al. "Lower bounds for non-convex stochastic optimization". In: *arXiv preprint arXiv:1912.02365* (2019).

[29] Angelia Nedić and Alex Olshevsky. "Stochastic gradient-push for strongly convex functions on time-varying directed graphs". In: *IEEE Transactions on Automatic Control* 61.12 (2016), pp. 3936–3947.

[30] Ran Xin, Usman A Khan, and Soummya Kar. "A near-optimal stochastic gradient method for decentralized non-convex finite-sum optimization". In: *arXiv preprint arXiv:2008.07428* (2020).

[31] Haoran Sun, Songtao Lu, and Mingyi Hong. "Improving the sample and communication complexity for decentralized non-convex optimization: Joint gradient estimation and tracking". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 9217–9228.

[32] Yann LeCun et al. "Backpropagation applied to handwritten zip code recognition". In: *Neural computation* 1.4 (1989), pp. 541–551.

[33] Anestis Antoniadis, Irène Gijbels, and Mila Nikolova. "Penalized likelihood regression for generalized linear models with non-quadratic penalties". In: *Annals of the Institute of Statistical Mathematics* 63.3 (2011), pp. 585–615.

[34] Zaid J Towfic and Ali H Sayed. "Stability and performance limits of adaptive primal-dual networks". In: *IEEE Transactions on Signal Processing* 63.11 (2015), pp. 2888–2903.

[35] Kun Yuan, Wei Xu, and Qing Ling. "Can primal methods outperform primal-dual methods in decentralized dynamic optimization?" In: *IEEE Transactions on Signal Processing* 68 (2020), pp. 4466–4480.

[36] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (2012), pp. 1097–1105.

[37] Andrew Maas et al. "Learning word vectors for sentiment analysis". In: *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*. 2011, pp. 142–150.

[38] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[39] Hamed Karimi, Julie Nutini, and Mark Schmidt. "Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition". In: *Joint European conference on machine learning and knowledge discovery in databases*. Springer. 2016, pp. 795–811.

Yiyue Chen received the B.S. degree from Wuhan University, China, in 2018, and the M.S.E. degree from the University of Texas at Austin in 2020. She received her Ph.D. degree in electrical and computer engineering from the University of Texas at Austin. Her research interests include distributed machine learning and optimization.

Abolfazl Hashemi (Member, IEEE) received the B.Sc. degree in electrical engineering from the Sharif University of Technology, Tehran, Iran, in July 2014, and the M.S.E. and Ph.D. degrees in electrical and computer engineering from the University of Texas at Austin, Austin, TX, USA, in May 2016 and August 2020, respectively. From August 2020 to August 2021, he was a Postdoctoral Scholar with the Oden Institute for Computational Engineering and Sciences, the University of Texas at Austin. Since August 2021, he has been an Assistant Professor with the Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA. Abolfazl was the 2019 Schmidt Science Fellows Award nominee from UT Austin and was the recipient of the Iranian National Elite Foundation Fellowship and a Best Student Paper Award finalist at the 2018 American Control Conference. His research interests include optimization for machine learning, signal processing, and control.

Haris Vikalo received the B.S. degree from the University of Zagreb, Croatia, in 1995, the M.S. degree from Lehigh University in 1997, and the Ph.D. degree from Stanford University in 2003, all in electrical engineering. He held a short-term appointment at Bell Laboratories, Murray Hill, NJ, in the summer of 1999. From January 2003 to July 2003 he was a Postdoctoral Researcher, and from July 2003 to August 2007 he was an Associate Scientist at the California Institute of Technology. Prof. Vikalo has been with the Department of Electrical and Computer Engineering, the University of Texas at Austin, since September 2007. He is a recipient of the 2009 National Science Foundation Career Award. His research interests include signal processing, machine learning, communications and bioinformatics.