Decentralized Composite Optimization in Stochastic Networks: A Dual Averaging Approach With Linear Convergence

Changxin Liu, Member, IEEE, Zirui Zhou, Jian Pei, Fellow, IEEE, Yong Zhang, and Yang Shi, Fellow, IEEE

Abstract—Decentralized optimization, particularly the class of decentralized composite convex optimization (DCCO) problems, has found many applications. Due to ubiquitous communication congestion and random dropouts in practice, it is highly desirable to design decentralized algorithms that can handle stochastic communication networks. However, most existing algorithms for DCCO only work in networks that are deterministically connected during bounded communication round(s), and therefore, cannot be extended to stochastic networks. In this article, we propose a new decentralized dual averaging (DDA) algorithm that can solve DCCO in stochastic networks. Under a rather mild condition on stochastic networks, we show that the proposed algorithm attains global linear convergence if each local objective function is strongly convex. Our algorithm substantially improves the existing DDA-type algorithms as the latter were only known to converge sublinearly prior to our work. The key to achieving the improved rate is the design of a novel dynamic averaging consensus protocol for DDA, which intuitively leads to more accurate local estimates of the global dual variable. To the best of our knowledge, this is the first linearly convergent DDA-type decentralized algorithm and also the first algorithm that attains global linear convergence for solving DCCO in stochastic networks. Numerical results are also presented to support our design and analysis.

Index Terms—Consensus control, composite optimization, distributed optimization, dual averaging, multi-agent systems.

I. INTRODUCTION

Consider a group of $n$ agents (e.g., processors and machines), each of which has its own objective function. They are connected via a bidirectional communication network and aim to cooperatively solve the following convex composite optimization problem in a decentralized manner:

$$
\min_{x \in \mathbb{R}^m} \left\{ F(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x) + h(x) \right\}
$$

(1)

where $f_i$ is the local smooth objective function of the agent $i$, and $h$ is a nonsmooth regularization term that is shared across all the agents. Problem (1) is referred to as decentralized convex composite optimization [1], [2] and finds broad applications in optimal control of multiagent systems [3], resource allocation [4], and large-scale machine learning [5], just to name a few [6].

In this work, we focus on solving Problem (1) when the communication network is stochastic. There are many practical reasons that promote the consideration of stochastic communication networks. Indeed, communication in real networks is usually subject to congestion, errors, and random dropouts, which is typically modeled as a stochastic process. Besides, stochastic networks are useful for proactively reducing communication cost. For instance, the gossip protocol [7] and Bernoulli protocol [8], which randomly choose a subset of communication links from an underlying dense graph in each iteration, have been widely regarded as effective strategies to avoid high communication cost and network congestion. Therefore, it is highly desirable to develop decentralized algorithms that solve Problem (1) over stochastic communication networks and attain a favorable convergence rate.

Over the past decade, many algorithms have been proposed for solving Problem (1). Some of them exploit the composite structure in (1) and attain global linear convergence if Problem (1) is strongly convex (see, e.g., [2] and [9]), which is the fastest rate of convergence that one can expect from a first-order decentralized algorithm. However, such linear convergence results are limited to time-invariant communication networks, because the design of these algorithms inherently requires knowledge of network topology a priori. Indeed, these algorithms are typically developed upon leveraging centralized primal-dual optimization paradigms, such as the alternating direction method of multipliers (ADMM) [10], to solve the following problem that is equivalent to (1):

$$
\min_{x_1, \ldots, x_n \in \mathbb{R}^m} \frac{1}{n} \sum_{i=1}^{n} (f_i(x_i) + h(x_i)) \quad \text{s.t.} \quad (\mathcal{L} \otimes I)x = 0
$$

(2)
where \( \mathbf{x} = [x_1^T, \ldots, x_m^T]^T \), \( \otimes \) denotes the Kronecker product, \( I \) is an identity matrix of size \( m \times m \), and \( \mathcal{L} \) denotes the graph Laplacian associated with the communication network. Since \( \mathcal{L} \) needs to be explicitly given in formulation (2), these algorithms and their associated linear convergence results cannot be extended to stochastic communication networks, where the network topology is time varying and random.

Among the existing decentralized optimization methods, the decentralized dual averaging (DDA) algorithm proposed by [11] and its later extensions [12], [13], [14] have been recognized as a powerful framework that can handle stochastic networks. However, the convergence rates of existing DDA-type algorithms are rather slow. In fact, even for decentralized convex smooth optimization in time-invariant networks, which is deemed to be much simpler than Problem (1) in stochastic networks, these algorithms were only known to converge sublinearly. Specifically, existing DDA-type algorithms, when applied to Problem (1), only attain an \( O(1/\sqrt{t}) \) sublinear rate of convergence. For the special case of Problem (1) with \( h \equiv 0 \), the authors in [15] recently showed that the convergence rate can be improved to \( O(1/t) \). Nevertheless, it remains open whether a DDA-type algorithm can attain linear rate of convergence.

**Contribution:** In this article, we propose a new DDA algorithm that solves Problem (1) in stochastic networks. Under a rather mild condition on the stochastic network, we show that the proposed algorithm has an \( O(1/t) \) rate of convergence in the general case and a global linear rate of convergence if each local objective function is strongly convex. Our work contributes to the literature of decentralized optimization in the following two aspects.

1) We develop the first decentralized algorithm that attains global linear convergence for solving Problem (1) in stochastic networks. Existing linearly convergent decentralized algorithms for Problem (1) only work in networks that are deterministically connected during bounded communication rounds, and therefore, cannot be extended to stochastic networks. Our algorithm is based on a DDA framework that is fundamentally different from these algorithms.

2) Our algorithmic design and convergence analysis shed new light on DDA-type algorithms. Notably, it is the first DDA-type algorithm that attains linear convergence. Prior to our work, even for decentralized convex smooth optimization in time-invariant networks, existing DDA-type algorithms were only known to converge sublinearly. The key to achieving the improved rate is the design of a novel dynamic averaging consensus protocol for DDA, which intuitively leads to more accurate local estimates of the global dual variable.

**II. RELATED WORKS**

**A. Decentralized Algorithms for Problem (1) in Deterministic Networks**

Due to its broad applications, Problem (1) has received attention in the community of decentralized optimization for many years; see, e.g., [1] for an early attempt. It is only until recently that linearly convergent decentralized algorithms have been developed for solving Problem (1) in deterministic networks. For time-invariant networks, Alghunaim et al. [2] developed a decentralized proximal gradient method, where the diffusion step and the proximal step are designed differently from [1] such that not only the fixed point meets the global optimality condition but also linear convergence can be attained for strongly convex problems. Furthermore, the strategy was generalized as a unified framework for proximal gradient tracking in [16]. Latafat et al. [9] proposed a distributed algorithm based on the randomized block-coordinate proximal method, which exhibits an asymptotic linear convergence if the monotone operator associated with Problem (1) is metrically subregular (a much weaker condition than strong convexity). Very recently, Xu et al. [17] proposed a unified decentralized algorithmic framework based on the operator splitting theory, which attains linear convergence for the strongly convex case. For deterministic time-varying networks, the authors in [18] developed a linearly convergent decentralized optimization algorithm based on the gradient-tracking technique and elaborate objective surrogates. However, it still requires the network to be connected during bounded communication rounds, which is a worst-case assumption about network connectivity [19] and does not necessarily hold in stochastic networks. To summarize, existing linearly convergent decentralized algorithms for Problem (1) are only applicable to deterministic networks and cannot be extended to the stochastic networks, which motivates the new algorithm development and convergence analysis in this article.

**B. Decentralized Optimization in Stochastic Networks**

The study of decentralized algorithms over stochastic networks dates back to [19], who proposed a subgradient-based algorithm with diminishing step sizes. The DDA algorithm, which combines the dual averaging method [20] and consensus seeking, was reported in [11] and can handle stochastic networks with an \( O(1/\sqrt{t}) \) sublinear rate of convergence. The decentralized accelerated gradient algorithm with a random network model was proposed in [21], where an \( O(\log t) \) sublinear convergence rate is obtained for smooth problems. A decentralized ADMM algorithm was designed in [22], where a few nodes are randomly selected to perform local updates. Decentralized optimization with asynchronous local updates was considered in [23] and [24]. Later, the authors in [25] and [26] validated the use of a constant step size in decentralized gradient descent over stochastic networks, leading to a global linear rate of convergence for strongly convex and smooth problems. Recently, Koloskova et al. [27] developed a unified framework for decentralized stochastic gradient descent over stochastic networks. It is worth mentioning that the aforementioned studies either consider general nonsmooth problems or focus on smooth problems. In particular, they cannot exploit the composite structure of Problem (1), partially due to the technical difficulty caused by the so-called projection-consensus coupling [11] for methods integrating consensus-seeking and projected/proximal gradient descent.
III. PRELIMINARIES

A. Basic Setup

We consider the finite-sum optimization problem (1), in which \( h \) is a closed convex function with its domain, denoted by \( \text{dom}(h) \), being nonempty, and \( f_i \) satisfies the following assumptions for all \( i = 1, \ldots, n \). Typical choices of \( h \) include the elastic net regularization, i.e., \( h(x) = \lambda_1 \|x\|_1 + \lambda_2 \|x\|_2^2 \), \( \lambda_1, \lambda_2 \geq 0 \), and the indicator function of a closed convex set.

Assumption 1:

i) \( f_i \) is continuously differentiable on an open set that contains \( \text{dom}(h) \).

ii) \( f_i \) is (strongly) convex with modulus \( \mu \geq 0 \) on \( \text{dom}(h) \), i.e., for any \( x, y \in \text{dom}(h) \)

\[
f_i(x) - f_i(y) - \langle \nabla f_i(y), x - y \rangle \geq \frac{\mu}{2} \|x - y\|^2. \tag{3}
\]

iii) \( \nabla f_i \) is Lipschitz continuous on \( \text{dom}(h) \) with Lipschitz constant \( L > 0 \), i.e., for any \( x, y \in \text{dom}(h) \)

\[
\| \nabla f_i(x) - \nabla f_i(y) \| \leq L \|x - y\|. \tag{4}
\]

Throughout this article, we denote by \( x^* \) an optimal solution of Problem (1). Assumption 1 is standard in the study of decentralized optimization [1], [28]. It is worth noting that we allow \( \mu = 0 \) in Assumption 1(ii), which reduces to the general convex case.

B. Stochastic Communication Networks

We consider solving Problem (1) in a decentralized manner, that is, each agent \( i \) holds a local objective function \( F_i := f_i + h \) and a pair of agents can exchange information only if they are connected in the communication network. Similar to existing studies [1], [11], [29], [30], we use a doubly stochastic matrix \( P(t) \in [0, 1]^{n \times n} \) to encode the network topology and the weights of connected links at time \( t \). We focus on the fairly general setting of stochastic communication networks, i.e., \( P(t) \) is a random matrix for every \( t \). For the convergence of the proposed decentralized algorithm, we make the following assumption on \( P(t) \).

Assumption 2: For every \( t \geq 0 \), it holds the following:

i) the network is undirected;

ii) \( P(t) \mathbf{1} = 1 \) and \( \mathbf{1}^T P(t) = 1^T \), where \( \mathbf{1} \) denotes the all-one vector of dimensionality \( n \);

iii) \( P(t) \) is independent of the random events that occur up to time \( t - 1 \);

iv) there exists a constant \( \beta \in (0, 1) \) such that

\[
\sqrt{\rho \left( E_t \left[ P(t)^T P(t) \right] - \frac{11^T}{n} \right)} \leq \beta \tag{5}
\]

where \( \rho(\cdot) \) denotes the spectral radius and the expectation \( E_t[\cdot] \) is taken with respect to the distribution of \( P(t) \) at time \( t \).

Assumption 2 has been used for analyzing the convergence of a host of decentralized algorithms; see, e.g., [7], [25], and [27]. It is satisfied by numerous stochastic communication settings; we take the following two common settings as examples.

1) Randomized gossip: At every time \( t \) one communication link \((i, j)\) is sampled from an underlying graph \( \mathcal{G} \). Suppose that we take \( P(t) = I - \frac{1}{n} \mathbf{e}_i \mathbf{e}_j^T \), where \( I \) is the identity matrix and \( \mathbf{e}_i \in \mathbb{R}^n \) is a vector with 1 in the \( i \)th position and 0 otherwise. Then, it is known that Assumption 2 is satisfied provided that the underlying graph \( \mathcal{G} \) is connected; see, e.g., [7].

2) Bernoulli stochastic networks: Consider an underlying graph \( \mathcal{G} \), where the state (online or offline) of each link \((i, j)\) is a Bernoulli process with link probability \( w_{ij} \). Denote by \( L(t) \) the Laplacian at time \( t \), and set \( P(t) = I - L(t)/(2d) \), where \( d = \max_{i} d_i \) and \( d_i \) is the degree of node \( i \) in \( \mathcal{G} \). It can be verified that Assumption 2 holds when the second largest eigenvalue of Laplacian average \( \mathbb{Z} \) is strictly positive [8].

C. Dual Averaging Method

Our algorithm is based on the dual averaging method that was originally proposed in [20]. The dual averaging method originally proposed by [20] can be directly applied to solve Problem (1) in a centralized manner. In particular, let \( d \) be a strongly convex function with modulus 1 on \( \text{dom}(h) \) such that

\[
x(0) = \arg\min_{x \in \mathbb{R}^n} d(x) \in \text{dom}(h) \quad \text{and} \quad d(x(0)) = 0. \tag{6}
\]

Then, the dual averaging method starts with \( x(0) \) and iteratively generates \( \{x(t)\}_{t \geq 1} \) according to

\[
x(t) = \arg\min_{x \in \mathbb{R}^n} \left\{ \sum_{s=0}^{t-1} a_{t-s+1} \ell(x; x(s)) + d(x) \right\} \tag{7}
\]

where

\[
a_t = \frac{a}{1 - a \mu t}, \quad t = 1, 2, \ldots \tag{8}
\]

for some constant \( a > 0 \), \( \ell: \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R} \) is defined as

\[
\ell(y; z) := f(z) + \langle \nabla f(z), y - z \rangle + \frac{\mu}{2} \|y - z\|^2 + h(y) \tag{9}
\]

for any \( y, z \in \mathbb{R}^m \), and \( f = \frac{1}{n} \sum_{i=1}^n f_i \). It is worth noting that for the strongly convex case (i.e., \( \mu > 0 \)), the sequence \( \{a_t\}_{t \geq 1} \) is geometrically increasing; for the general convex case (i.e., \( \mu = 0 \)), the sequence \( \{a_t\}_{t \geq 1} \) equals the constant \( a \). Moreover, both (7) and (8) require the modulus \( \mu \) of strong convexity. In practice, one can use a lower bound of \( \mu \) or simply set \( \mu = 0 \) in (7) and (8) if no valid lower bound is available.

The following theorem summarizes the convergence property of the aforementioned dual averaging method, which is a direct extension of [31, Th. 3.2] to problems with nonsmooth regularization terms. A proof of Theorem 1 is provided in [32, Appendix F].

**Theorem 1:** When Assumption 1 is satisfied, let \( \{x(t)\}_{t \geq 0} \) be the sequence of iterates generated by the dual averaging method...
method (7), if $a \leq L^{-1}$, then

$$F(\tilde{x}^{(t)}) - F(x^t) \leq \frac{d(x^t)}{A_t}, \quad t = 1, 2, \ldots$$

where $A_t = \sum_{\tau=1}^t a_{\tau}$ and $\tilde{z}^{(t)} = A_t^{-1} \sum_{\tau=1}^t a_{\tau} x^{(\tau)}$. Moreover, the following estimates on $A_t^{-1}$ hold.

1. If $\mu > 0$, then $\frac{1}{A_t} \leq \frac{1}{a_0}\left(1 - \frac{\mu}{\gamma_a}\right)^t$.
2. If $\mu = 0$, then $\frac{1}{A_t} = \frac{1}{a_t}$.

**IV. ALGORITHM AND MAIN RESULTS**

From Theorem 1, one can observe that the dual averaging method, when applied to solve Problem (1) in a centralized manner, attains global linear convergence if Problem (1) is strongly convex. The existing dual-averaging-based decentralized algorithms, however, converge only sublinearly. In view of this, the following question arises naturally: can we develop a dual-averaging-based decentralized algorithm that can achieve the same order of convergence as its centralized counterpart, that is, linear convergence? In this section, we put an affirmative answer to this question by developing a new DDA algorithm that incorporates a novel dynamic averaging consensus protocol for each local update, which intuitively leads to more accurate local estimates of the global dual variable. We show that the new DDA, when applied to solve Problem (1) in stochastic networks, converges linearly if each local objective is strongly convex. Our algorithm design and convergence analysis shed new light on DDA-type algorithms, as it is the first DDA-type algorithm that can achieve linear convergence. Besides, it is also the first linearly convergent algorithm for solving Problem (1) in stochastic networks.

To motivate the design of our DDA method, we observe that by letting $A_t = \sum_{\tau=1}^t a_{\tau}$ and

$$z^{(t)} = \sum_{r=0}^{t-1} a_{r+1} \left(\frac{1}{n} \sum_{i=1}^n \nabla f_i(x^{(\tau)}) - \mu x^{(\tau)}\right)$$

the update rule (7) can be written as

$$x^{(t)} = \text{argmin}_{x \in \mathbb{R}^n} \left\{ \langle z^{(t)}, x \rangle + A_t \left(\frac{H}{2} \|x\|^2 + h(x)\right) + d(x) \right\}. \quad (10)$$

Thus, it is sensible for each agent to locally estimate the global dual variable $z^{(t)}$ to fulfill decentralization. To this end, we propose the following dynamic averaging consensus protocol:

$$z_i^{(t)} = \sum_{j=1}^n p_{ij}^{(t-1)} z_j^{(t-1)} + a_i s_i^{(t-1)} \quad (11a)$$

$$s_i^{(t)} = \sum_{j=1}^n p_{ij}^{(t-1)} s_j^{(t-1)} + \left(\nabla f_i(x_i^{(t-1)}) - \mu x_i^{(t-1)}\right) - \left(\nabla f_i(x_i^{(t-1)}) - \mu x_i^{(t-1)}\right) \quad (11b)$$

where $p_{ij}^{(t)}$ is the $(i, j)$th element in the mixing matrix $P^{(t)}$, $z_i^{(t)}$ is the $i$th agent’s local estimate of $z^{(t)}$ at time $t$, and $s_i^{(t)}$ is an auxiliary vector for reducing the consensus error. Equipped with these, each agent $i$ can perform a local computation to update its estimate of the global primal variable $x^{(t)}$ as

$$x_i^{(t)} = \text{argmin}_{x \in \mathbb{R}^n} \left\{ \langle z_i^{(t)}, x \rangle + A_t \left(\frac{H}{2} \|x\|^2 + h(x)\right) + d(x) \right\}. \quad (12)$$

We denote by $N_i^{(t)}$ the set of agents that are connected with the agent $i$ at time $t$. Then, the entire algorithm can be summarized in Algorithm 1.

**Algorithm 1:** The proposed DDA algorithm for Problem (1).

1. **Input:** $\mu \geq 0$, $a > 0$, $x^{(0)} \in \text{dom}(h)$ and a strongly convex function $d$ with modulus 1 on $\text{dom}(h)$ such that (6) holds.
2. **Initialize:** $a_0 = a$, $A_0 = 0$, $x_i^{(0)} = x^{(0)}$, $z_i^{(0)} = 0$, and $s_i^{(0)} = \nabla f_i(x^{(0)}) - \mu x^{(0)}$ for all $i = 1, \ldots, n$.
3. **for** $t = 1, 2, \ldots$ **do**
4. set $a_t = a_{t-1}/(1 - \mu)$ and $A_t = A_{t-1} + a_t$.
5. In parallel (for agent $i, i = 1, \ldots, n$):
6. collect $z_j^{(t-1)}$ and $s_j^{(t-1)}$ from all agents $j \in N_i^{(t-1)}$.
7. update $z_i^{(t)}$ and $s_i^{(t)}$ by (11).
8. compute $x_i^{(t)}$ by (12).
9. broadcast $z_i^{(t)}$ and $s_i^{(t)}$ to all agents $j \in N_i^{(t)}$.
10. **end for**

Our protocol (11) differs from the one used in the original DDA [11] in (11b), where the latter simply lets $s_i^{(t)} = \nabla f_i(x_i^{(t)})$ for all agents $i$. Our update in (11b) is a second-order dynamic averaging consensus protocol motivated by [33], and equips each agent $i$ with an $s_i^{(t)}$ that can track the global variable $\frac{1}{n} \sum_{i=1}^n (\nabla f_i(x_i^{(t)}) - \mu x_i^{(t)})$. Intuitively, (11) can lead to much more accurate local estimates $\{z_i^{(t)}\}_{i=1}^n$ of the global dual variable $z^{(t)}$. Moreover, as we will show in the convergence analysis, the novel update (11b) validates the use of geometrically increasing weights $\{a_t\}_{t \geq 0}$ in (11a). This contrasts with the use of decaying weights in other DDA-type algorithms and is key to achieving the linear convergence result.

Before proceeding, we make some remarks on Algorithm 1. First, Algorithm 1 provides a unified treatment for both general convex and strongly convex cases. In particular, if $\mu = 0$, we simply set $a_t = a$ and $A_t = a t$ for all $t$. Second, to satisfy the condition in (6), one can choose an arbitrary $x^{(0)} \in \text{dom}(h)$ and let $d(x) := d(x) - d(x^{(0)}) - \langle \nabla d(x^{(0)}), x - x^{(0)} \rangle$, where $d$ is any strongly convex function with modulus 1, e.g., $d(x) = \|x\|^2/2$. It is easy to verify that such $x^{(0)}$ and $d$ satisfy (6). Third, for agent $i$ with $N_i^{(t-1)} = \emptyset$, it will not perform Step 6. However, all the agents are required to compute $x_i^{(t)}$ according to Steps 7 and 8. Finally, similar to the standard dual averaging method, we assume that the subproblem (12) can be computed easily. This holds for a host of applications. For example, if we choose $d(x) = \|x - x^{(0)}\|^2/2$, then the subproblem (12) reduces to computing the proximal operator of $A_t h/(1 + \mu A_t)$, which admits a closed-form solution in many applications. Compared to ADMM-based methods [34], where typically a nontrivial dual problem is solved at each iteration, the proposed method
has lighter computational cost per step. When subproblem (12) cannot be computed efficiently, one may run another loop to compute an approximate solution, which is common in decentralized composite optimization.

Remark 1 (Intuition behind (11)): In the centralized dual averaging update (10), only \(z^{(t)}\) contains global information. Therefore, if \(z^{(t)}\) can be estimated sufficiently accurate by the agents, they can solve (1) in a decentralized way. We follow the idea in [33] that the second-order dynamic averaging consensus can be used to estimate the average of local signals whose second-order differences are relatively bounded. Particularly, observe that

\[
z_i^{(t)} = \sum_{\tau=0}^{t-1} a_{t+1} \left( \nabla f_i(x_i^{(t)}) - \mu x_i^{(t)} \right).
\]

Take \(\nabla f_i(x_i^{(t)}) - \mu x_i^{(t)} - (\nabla f_i(x_i^{(t-1)}) - \mu x_i^{(t-1)})\) as the second-order difference of \(z_i^{(t)}\). By (4), one obtains its upper bound

\[
\|\nabla f_i(x_i^{(t)}) - \mu x_i^{(t)} - (\nabla f_i(x_i^{(t-1)}) - \mu x_i^{(t-1)})\| \\
\leq (L + \mu)\|x_i^{(t)} - x_i^{(t-1)}\|\|y_i\| \leq 1, \ldots, n.
\]

Therefore, the dynamic averaging consensus scheme can be used to estimate \(n^{-1} \sum_{i=1}^{n} z_i^{(t)}\), which is a good approximation of \(z^{(t)}\). This motivates our update formulas (11).

Remark 2: (Comparison with gradient-tracking methods [25, 35]) Recall the update in [25] and [35] as follows:

\[
x_i^{(t)} = \sum_{j=1}^{n} \theta_i^{(t-1)} \left( j \right) (x_j^{(t-1)} - a s_j^{(t-1)})
\]

\[
s_i^{(t)} = \sum_{j=1}^{n} \theta_i^{(t-1)} s_j^{(t-1)} + \nabla f_i \left( t_i^{(t)} \right) - \nabla f_i \left( t_i^{(t-1)} \right).
\]

It contains three key differences from our update in (11).

1. Equation (13) updates \(s_i^{(t)}\) by estimating the average of local gradients \(n^{-1} \sum_{i=1}^{n} \nabla f_i(x_i^{(t-1)})\), but (11) updates \(s_i^{(t)}\) by estimating the term \(n^{-1} \sum_{i=1}^{n} (\nabla f_i(x_i^{(t)}) - \mu x_i^{(t)})\).

2. Equation (13) weights \(s_i^{(t)}\) with constant \(a\), but (11) weights \(s_i^{(t)}\) with geometrically increasing \(\{a_t\}_{t \in \mathbb{N}}\) that is motivated by the dual averaging method.

3. In the proposed DDA method, another proximal operator is performed over \(s_i^{(t)}\) to get \(x_i^{(t)}\). The gradient evaluated over \(x_i^{(t)}\) is then used to update \(s_i^{(t)}\). While (13) does not accommodate proximal operators.

The rest of this section presents the convergence results of Algorithm 1. To proceed, we denote

\[
M = \begin{bmatrix}
\frac{\beta}{a(L+\mu)} & \frac{\beta}{a(L+\mu)} \\
\frac{1}{a} & \frac{1}{a}
\end{bmatrix}
\]

where \(L\) and \(\mu\) are given in Assumption 1, \(\beta \in (0, 1)\) is defined in Assumption 2, and \(a\) is an input of Algorithm 1. Matrix \(M\) is the key to our convergence analysis as it defines the dynamics of the iterates generated by Algorithm 1. Let \(\rho(M)\) be the spectral radius of \(M\). To facilitate the presentation of our convergence analysis, we define

\[
\nu := \rho(M) \sqrt{1 - a\mu}
\]

\[
\eta := (1 - a\mu)(1 - \nu)^2
\]

\[
\theta := (1 - a\mu)(1 - \nu^2).
\]

The following result on \(\nu, \eta,\) and \(\theta\) is fundamental to our convergence analysis whose proof can be found in Appendix D-A.

Lemma 1: The value of \(\nu\) monotonically increases with \(a\) if \(a \in (0, 1/\mu)\). Moreover, if

\[
\frac{1}{\alpha} > \frac{\beta(2L + 3\mu)}{(1 - \beta)^2} + \mu
\]

then \(\nu < 1\). Consequently, \(\eta\) and \(\theta\) are both positive and monotonically decrease with \(a\) if (16) is satisfied.

Equipped with Lemma 1, we are ready to present the main results of this article, which pertain to the convergence property of Algorithm 1. Similar to some existing works [11], we first present the convergence property of an auxiliary sequence \(\{y^{(t)}\}_{t \geq 0}\), which then immediately implies the convergence property of the sequence \(\{x_i^{(t)}\}_{t = 1, \ldots, n}\) generated by Algorithm 1. In particular, we define

\[
y^{(t)} = \arg\min_{x \in \mathbb{R}^n} \left\{ \langle \sigma(x), x \rangle + A_t \left( \frac{\beta}{2} \|x\|^2 + h(x) + d(x) \right) \right\}
\]

where \(y^{(0)} = x^{(0)}, \sigma^{(t)} = \frac{1}{n} \sum_{i=1}^{n} z_i^{(t)}\), and \(\{z_i^{(t)} : i = 1, \ldots, n\}_{t \geq 0}\) are generated by Algorithm 1.

Theorem 2: When Assumptions 1 and 2 are satisfied, the constant \(a\) in Algorithm 1 satisfies (16), and

\[
\gamma := \frac{1}{\alpha} > 2L + \frac{4L - 2\mu}{\eta} > 0
\]

where \(\eta\) is defined in (15), then, for all \(t \geq 1\), it holds that

\[
E[F(y^{(t)})] - F(x^*) \leq \frac{C}{A_t}
\]

where \(\hat{y}^{(t)} = A_t^{-1} \sum_{t=1}^{t} x_t^{(t)}\) with \(y^{(t)}\) defined in (17)

\[
C := d(x^*) + \frac{a(2L - \mu)\sigma^2}{n\theta(L + \mu)^2} > 0,
\]

and \(\sigma^2\) is the variance of local gradients at \(t = 0\), i.e.,

\[
\sigma^2 = \frac{n}{n} \sum_{i=1}^{n} \left\| \nabla f_i(x^{(0)}) - \frac{1}{n} \sum_{j=1}^{n} \nabla f_j(x^{(0)}) \right\|^2.
\]

Moreover, for all \(t \geq 1\) and \(i = 1, \ldots, n\), we have

\[
E[\|\hat{x}_i^{(t)} - \hat{y}^{(t)}\|^2] \leq \frac{D}{A_t}
\]

where \(\hat{x}_i^{(t)} = A_t^{-1} \sum_{t=1}^{t} x_t^{(t)}\) and \(D = \frac{4nC}{\eta} + \frac{2a\sigma^2}{n(L + \mu)^2} > 0\).

The proof of Theorem 2 is postponed to Appendix B. Theorem 2 can be regarded as a decentralized counterpart of Theorem 1. Due to the presence of consensus error in the decentralized setting, Theorem 2 requires a more delicate choice of \(a\) for convergence. It is shown in [32, Appendix E] that there exists
an \( a > 0 \) such that any \( a \in (0, \tilde{a}) \) satisfies the conditions in Theorem 2. Moreover, \( \tilde{a} \) is roughly in the order \( O((1 - \beta)^2 / L) \).

As a consequence of Theorem 2, we show in Corollary 1 that Algorithm 1 attains global linear convergence if \( \mu > 0 \). Its proof is given in Appendix C-A.

**Corollary 1:** Suppose that the premise of Theorem 2 holds. If \( \mu > 0 \), then for all \( t \geq 1 \) and \( i = 1, \ldots, n \), we have

\[
\mathbb{E}[\|\tilde{x}_i^{(t)} - x^*\|^2] \leq \frac{2}{a} \left( \frac{2C}{\mu} + D \right) \left( 1 - a\mu \right)^t
\]

where \( \tilde{x}_i^{(t)} = A_i^{-1} \sum_{r=1}^{t} a_r x_i^{(r)} \), and \( C \) and \( D \) are positive constants given in Theorem 2.

To the best of our knowledge, Corollary 1 provides the first linear convergence result for solving Problem (1) in stochastic networks. It is also the first linear convergence result for any DDA-type algorithms.

**Remark 3:** Corollary 1 requires \( \mu \) to be uniform for all agents. From a technical perspective, this is necessarily made to ensure that (38) in the proof of Theorem 2 and Lemma 5 remain valid. The authors in [16] proved that with agent specific non-smooth regularization, linear convergence cannot be achieved (in the worst case) for decentralized composite optimization even in time-invariant networks.

As Corollary 1 also holds when Algorithm 1 is applied to solving Problem (1) with \( h \equiv 0 \) in time-invariant networks, it would be interesting to compare our linear convergence result with those of decentralized algorithms that also converge linearly in this special case. Based on the aforementioned remark on \( a \), one can observe that the rate of linear convergence in Corollary 1 is roughly \( O((1 - \rho)^t) \), where \( \rho = \Theta((1 - \beta)^2 / \kappa) \). This rate is suboptimal for decentralized convex smooth optimization in time-invariant networks, where a better rate \( \rho = \Theta((1 - \beta) / \kappa) \) is achieved by, for example, [17]. This is mainly due to the consensus-based gradient-tracking mechanism. DIGing [28], Harnessing [35], and AugDGM [25] are based on similar strategies, for which the convergence rate is \( \Theta((1 - \beta)^2 / \kappa) \). Indeed, under the algorithmic framework based on operator splitting [17], the gradient-tracking mechanism essentially leads to a contraction matrix \( I - (I - P)^2 \) on the recursion. The convergence rates of decentralized primal-dual algorithms, e.g., EXTRA [30] and NIDS [36], typically have better dependence on the network topology, that is, \( \Theta((1 - \beta) / \kappa) \). However, they are limited to time-invariant networks to the best of our knowledge.

For the case \( \mu = 0 \), Theorem 2 implies that Algorithm 1 has a global \( O(1/t) \) rate of convergence. In particular, we have the following corollary whose proof is presented in Appendix C-B.

**Corollary 2:** Suppose that the premise of Theorem 2 holds. If \( \mu = 0 \), then for all \( t \geq 1 \) and \( i = 1, \ldots, n \), we have

\[
\mathbb{E}[F(\tilde{y}_i^{(t)})] - F(x^*) \leq \frac{C}{at}
\]

\[
\mathbb{E}[\|\tilde{x}_i^{(t)} - \tilde{y}_i^{(t)}\|^2] \leq \frac{D}{at}
\]

where \( \tilde{y}_i^{(t)} = \frac{1}{t} \sum_{r=1}^{t} y_i^{(r)} \), \( \tilde{x}_i^{(t)} = \frac{1}{t} \sum_{r=1}^{t} x_i^{(r)} \), and \( C \) and \( D \) are positive constants given in Theorem 2. In addition, if \( h \equiv 0 \) in Problem (1), \( d(x) = \|x\|^2 / 2 \), and

\[
\frac{1}{a} > 2L \max \left\{ \frac{\beta}{(1 - \beta)^2}, \frac{6}{(1 - \nu)^2} \right\}
\]

where \( \beta \) and \( \nu \) are given in (5) and (15), respectively, then we further have

\[
\mathbb{E}[F(\tilde{x}_i^{(t)})] - F(x^*) \leq \frac{1}{t} \left( \frac{n\|x^*\|^2}{2a} + \frac{6\sigma^2}{L(1 - \nu^2)} \right). \tag{25}
\]

Similar to some existing works (e.g., [37]), we can only ensure the \( O(1/t) \) rate for the objective value at the auxiliary sequence \( \{\tilde{y}_i^{(t)}\}_{t \geq 1} \) and the distance of each agent’s local estimate \( \tilde{x}_i^{(t)} \) to \( \tilde{y}_i^{(t)} \) when \( h \neq 0 \); see (22) and (23) respectively. The major difficulty is that we cannot derive \( \frac{1}{n} \sum_{i=1}^{n} x_i^{(t)} = \tilde{y}_i^{(t)} \) when \( h \neq 0 \), which prevents us from getting (43). It remains open whether the \( O(1/t) \) rate for the objective value at \( \{\tilde{x}_i^{(t)}\}_{t \geq 1} \), as in (25), can be established when \( h \neq 0 \) without additional assumptions. We leave it as future work.

**V. NUMERICAL EXPERIMENTS**

For the experiments, we consider the decentralized LASSO problem [37] and the decentralized sparse logistic regression problem [2]. We present numerical results of Algorithm 1 (named as DDA below), and compare it with the following algorithms.

1. **Proximal gradient exact first-order algorithm (PG-EXTRA) in [1]**
   
   \[
   z^{(t)} = z^{(t-1)} - x^{(t-1)} + \bar{P} (2x^{(t-1)} - x^{(t-2)}) - a(\nabla^{(t-1)} - \nabla^{(t-2)})
   \]
   
   \[
   x^{(t)} = \text{Prox}_h^\delta(z^{(t)})
   \]

   where \( \bar{P} = \frac{(I + P) \otimes \mathbf{1}_d}{2} \), \( h(x) = \sum_{i=1}^{n} h_i(x_i) \), and \( \text{Prox}_h^\delta(z) := \arg\min_{z \in \mathbb{R}^m} h(z) + \frac{1}{2\delta} \|z - z^*\|^2 \).

2. **Proximal primal-dual diffusion (P2D2) algorithm in [2]**
   
   \[
   z^{(t)} = (I - aB)x^{(t-1)} + (I - B)(x^{(t-1)} - \nabla^{(t-1)}) - a(\nabla^{(t-1)} - \nabla^{(t-2)})
   \]
   
   \[
   x^{(t)} = \text{Prox}_h^\delta(z^{(t)})
   \]

   where \( B = (I - P) \otimes \mathbf{1}_d \).

3. **Distributed subgradient method (DSM) in [19]**
   
   \[
   x^{(t)} = P^{(t-1)} x^{(t-1)} - a_{t-1}x^{(t-1)}
   \]

   where \( r^{(t)} \in \partial F(x^{(t)}) \) and \( F(x) = \sum_{i=1}^{n} F_i(x_i) \).

4. **Conventional DDA (named as C-DDA) in [38]**
   
   \[
   z^{(t)} = P^{(t-1)} x^{(t-1)} + r^{(t-1)}
   \]
   
   \[
   x^{(t)} = \arg\min_{x \in \mathbb{R}^m} \left\{ a_{t-1} (z^{(t)}, x) + d(x) \right\}
   \]

   where \( d(x) = \sum_{i=1}^{n} d_i(x_i) \).

We note that when applied to solve Problem (1) in stochastic networks, PG-EXTRA and P2D2 have no convergence guarantees and DSM and C-DDA have sublinear convergence in theory.
A. Strongly Convex Problems

The aforementioned algorithms are applied to the following problem:

$$\min_{x \in \mathbb{R}^m} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \|b_i - C_i x\|^2, \quad \text{s.t.} \quad \|x\|_1 \leq R$$

where $R > 0$ is a constant, and $(C_i, b_i)$ represents the data tuple available to the agent $i$ with $C_i \in \mathbb{R}^{60 \times 50}$ and $b_i \in \mathbb{R}^{60}$. The data are randomly generated according to the setting in [36]. First, a sparse signal $x^* \in \mathbb{R}^{50}$ is randomly generated, where the probability for each element being nonzero is 0.25. Then, each $C_i$ is randomly generated, and then, normalized such that $\|b_i - C_i x^*\|_1$ is produced based on $b_i = C_i x^* + \epsilon_i$, where $\epsilon_i$ is a random noise vector.

We consider two common configurations of stochastic communication networks. The first one is Bernoulli networks [8], where a fixed graph is first generated and at any time $t$, each edge of the fixed graph is sampled with probability $\iota \in (0, 1)$, which results in a random subgraph of the fixed graph. In our experiment, we generate a fixed graph in the same way as [34], where the sparsity parameter $\xi$, i.e., the ratio between the number of edges in the generated fixed graph and the number of edges in the complete graph, is chosen to be 0.2. Based on each fixed graph, we generate two Bernoulli networks with $\iota$ set to be 0.05 and 0.1, respectively. The second one is randomized gossip networks [7], where only a single edge of a fixed graph is sampled at any time $t$. In particular, the probability to sample the link $(i, j)$ is set as $\frac{1}{n(i+|\mathcal{N}_i|)}$, with $|\mathcal{N}_i|$ representing the number of neighbors of $i$ in the supergraph at every time $t$. In our experiment, we consider cycle graph, 2-D grid, and complete graph as the fixed graphs for generating randomized gossip networks.

For all the tested algorithms, we evaluate their performance in terms of the relative square error defined by

$$\sum_{i=1}^{n} \frac{\|x^* - x^i\|^2}{\sum_{i=1}^{n} \|x^i\|^2},$$

where $x^*$ is identified by applying the centralized proximal gradient method [39] to Problem (26) such that the norm of the difference of two consecutive iterates is less than $10^{-14}$. The algorithm by [40] is used to perform projection onto $l_1$-norm ball. All the algorithms are initialized with $x^i(0) = 0$ for all $i$. The parameters for each algorithm are chosen in the following way. For DDA and C-DDA, we employ $d(x) = \|x\|^2/2$. We choose $\alpha = 0.5$ in P2D2 and set $\alpha_i = 1/\sqrt{T} + 1$ for C-DDA. For the two groups of Bernoulli networks, we set the $a$ in DDA to be 0.1 and set 0.1 for the step sizes in P2D2 and PG-EXTRA. For randomized gossip, we use a $a = 0.1$ for DDA, and set $10^{-4}$ for the step sizes in P2D2 and PG-EXTRA. Since DSM cannot be applied to constrained problems, it is not considered in this setting.

The simulation results are plotted in Fig. 1. In particular, the performance on Bernoulli networks and randomized gossip networks is presented in the first and the second column of Fig. 1, respectively. In the first column, the bottom plot demonstrates the performance in the time-invariant network that is used for generating Bernoulli networks. Although P2D2 and PG-EXTRA demonstrate a similar performance with DDA on time-invariant networks, they do not converge to the minimizer when applied to stochastic networks. In line with our theoretical results, DDA linearly converges and outperforms C-DDA in all the network configurations.

B. General Convex Problems

The following decentralized sparse logistic regression problem is considered

$$\min_{x \in \mathbb{R}^m} \frac{1}{n} \sum_{i=1}^{n} f_i(x) + \phi \|x\|_1$$

where

$$f_i(x) = \frac{1}{m_i} \sum_{j=1}^{m_i} \ln(1 + \exp(-y_j^i M_j^T x))$$

and $(M_j^i, y_j^i)_{j=1}^{m_i}$ are data samples private to agent $i$. In our experiment, we set $\phi = 0.001$, and use Spambase dataset in the UCI Machine Learning Repository [41] to generate our problem instance. In particular, we extract 3000 out of the total 4601 samples in the original dataset and evenly distribute them to the $n = 30$ agents, i.e., $m_i = 100$ for all $i$.

Two types of stochastic communication networks are considered. For Bernoulli networks, we generate a fixed graph with the sparsity parameter 0.4. Based on it, we construct two Bernoulli networks by setting $\iota = 0.1$ and $\iota = 0.2$, respectively. In the
second setting, we also consider cycle graph, 2-D grid, and complete graph as the fixed graphs for generating randomized gossip networks. We identify $x^{\ast}$ by using the centralized proximal gradient method, where the stopping criterion is set as the norm of the difference of two consecutive iterates smaller than $10^{-14}$. The performance of all the tested algorithms is evaluated in terms of the suboptimality defined by $F(n^{-1}\sum_{i=1}^{n}x_{i}^{(t)}) - F(x^{\ast})$. All the algorithms are initialized with $x_{i}^{(0)} = 0$ for all agents $i$. The parameters of each algorithm are chosen properly to reflect their performance. For DDA and C-DDA, we simply choose $\alpha = 0.5$ in P2D2 and set $\alpha = 1/\sqrt{t + 1}$ for C-DDA and DSM. For the Bernoulli networks, i.e., those sampled from a supergraph with sparsity parameter 0.4 (first column of Fig. 2), we use the same $\alpha = 0.2$ for DDA, P2D2, and PG-EXTRA. For randomized gossip, we use $\alpha = 0.05$ for DDA, and set $10^{-4}$ for the step sizes in P2D2 and PG-EXTRA. We note that choosing a smaller step size in P2D2 and PG-EXTRA generally makes them more stabilizing. In fact, a larger step size will result in even worse behavior of these two methods in randomized gossip networks.

The simulation results are plotted in Fig. 2. Specifically, the first column of Fig. 2 presents the performance on Bernoulli networks and the second column shows the performance on randomized gossip networks. We note that the last plot of the first column is for the time-invariant network, which is the fixed graph for generating the Bernoulli networks in the first column. One can observe that our DDA is substantially faster than DSM and C-DDA in all the network settings, which supports our theoretical development. In addition, while P2D2 and PG-EXTRA perform very similar to DDA on time-invariant networks, they both diverge when applied to stochastic networks. This suggests that decentralized algorithms that are designed for time-invariant networks may not work effectively in stochastic networks.

To summarize, the simulation results confirm our theoretical findings and demonstrate the superior performance of the proposed Algorithm 1 on both time-invariant and stochastic networks.

VI. CONCLUSION AND FUTURE WORK

In this article, we proposed a new decentralized algorithm for solving Problem (1) in stochastic networks. The proposed algorithm, based on the framework of dual averaging method, is facilitated by designing a novel dynamic averaging consensus protocol. To the best of our knowledge, this is the first linearly convergent DDA-type decentralized algorithm and also the first algorithm that attains global linear convergence for solving Problem (1) in stochastic networks.

As we remarked after Corollary 1, it remains open whether Algorithm 1 can be further improved such that it achieves the optimal rate of linear convergence when applied to the special setting of Problem (1). Besides, it is unknown if $O(1/t)$ convergence rate of the objective error at the local estimates can be established for the general convex case. Another practical issue about decentralized optimization is the privacy risk due to information exchange among multiple agents. We leave them as future research.

Appendix A

Overview of the Appendix and Preliminaries

In the Appendix, we begin with some notation to streamline the presentation. Then, we present the proof of Theorem 2 and the proofs of Corollaries 1 and 2 in Appendixes B and C, respectively. The proofs of supporting lemmas are postponed to Appendix D.

First, we introduce the following notation:

\[
\begin{align*}
\bar{x}^{(t)} &= \begin{bmatrix} x_{1}^{(t)} \\ \vdots \\ x_{n}^{(t)} \end{bmatrix}, & s^{(t)} &= \begin{bmatrix} s_{1}^{(t)} \\ \vdots \\ s_{n}^{(t)} \end{bmatrix}, & \bar{z}^{(t)} &= \begin{bmatrix} z_{1}^{(t)} \\ \vdots \\ z_{n}^{(t)} \end{bmatrix} \\
\nabla^{(t)} &= \begin{bmatrix} \nabla f_{1}(x_{1}^{(t)}) \\ \vdots \\ \nabla f_{n}(x_{n}^{(t)}) \end{bmatrix}, & \bar{y}^{(t)} &= \begin{bmatrix} y_{1}^{(t)} \\ \vdots \\ y_{n}^{(t)} \end{bmatrix}, & \bar{z}^{(t)} &= \frac{1}{n} \sum_{i=1}^{n} z_{i}^{(t)} \\
\bar{x}^{(t)} &= \frac{1}{n} \sum_{i=1}^{n} x_{i}^{(t)}, & \bar{s}^{(t)} &= \frac{1}{n} \sum_{i=1}^{n} \nabla f_{i}(x_{i}^{(t)}), & \bar{x}^{(t)} &= \frac{1}{n} \sum_{i=1}^{n} s_{i}^{(t)} \\
\bar{s}^{(t)} &= \bar{s}^{(t)} - 1 \otimes \bar{s}^{(t)}, & \bar{z}^{(t)} &= \bar{z}^{(t)} - 1 \otimes \bar{s}^{(t)} \\
\end{align*}
\]

(27)
\[ \Delta x^{(t-1)} = x^{(t-1)} - x^{(t-1)}, \quad \Delta y^{(t-1)} = y^{(t)} - y^{(t-1)} \]  
(30)  
\[ \Delta x^{(t-1)} = x^{(t-1)} - x^{(t-1)}, \quad \Delta y^{(t-1)} = y^{(t)} - y^{(t-1)} \]  
(31)

where \( 1 \) is an all-one column vector of dimension \( n \). We remark that bold lowercase letters represent a vector of dimension \( m \times n \), while normal lowercase letters represent a vector of dimension \( m \). Equipped with these notation, we can rewrite the update rule (11) in the following compact form:

\[
z^{(t)} = P^{(t-1)} \left( y^{(t-1)} + a_t s^{(t-1)} \right)  
(32a)
\]

\[
s^{(t)} = P^{(t-1)} s^{(t-1)} + \nabla y - \nabla x^{(t-1)} - \mu \Delta x^{(t-1)}  
(32b)
\]

where \( P^{(t)} = P^{(t)} \otimes I \) with \( I \) being an identity matrix of size \( n \times n \).

For a real-valued random vector \( x \), we define

\[ \|x\|_2 = \sqrt{\mathbb{E} \|x\|^2}. \]  
(33)

Accordingly, for a square random matrix \( W \), we define \( \|W\|_2 = \sup_{\|x\|=1} \|Wx\|_2 \). Given two real-valued random vectors \( x, y \), the Minkowski inequality [42] states that

\[ \|x + y\|_2 \leq \|x\|_2 + \|y\|_2. \]  
(34)

Finally, it is known that Assumption 1(iii) implies that

\[ f_i(x) - f_i(y) - \langle \nabla f_i(y), x - y \rangle \leq \frac{L}{2} \|x - y\|^2 \]  
(35)

for any \( x, y \in \text{dom}(h) \).

**APPENDIX B**

**PROOF OF THEOREM 2**

In this section, we provide the proof of Theorem 2.

To start, we show the following result that quantifies the deviation between the local estimates \( \{x_i^{(t)}\}_{t \geq 0} \) and the auxiliary sequence \( \{y_i^{(t)}\}_{t \geq 0} \).

**Lemma 2:** Suppose that \( a \) satisfies (16). Then, for all \( t \geq 0 \), it holds that

\[
\sum_{\tau=0}^{t} a_{\tau+1} \mathbb{E} \|x^{(\tau)} - y^{(\tau)}\|^2 \leq \frac{2}{\eta} \sum_{\tau=0}^{t-1} a_{\tau+1} \mathbb{E} \|\Delta y^{(\tau)}\|^2 + \frac{2a_0 \sigma^2}{\theta (L + \mu)^2} \]  
(36)

where \( \sigma \) is defined in Theorem 2, \( \eta \) and \( \theta \) are given in (15), and both \( \eta \) and \( \theta \) are positive due to (16) and Lemma 1.

**Proof of Lemma 2:** The proof is postponed to Appendix 4-B. \( \blacksquare \)

Lemma 2 states that if \( a \) satisfies (16), then the accumulative deviation between \( y^{(t)} \) and \( x^{(t)} \) admits an upper bound constituted by the successive change of \( y^{(t)} \) plus a constant.

Next, we present the following lemma that pertains to a descent-like property of Algorithm 1.

**Lemma 3:** For all \( t \geq 1 \), it holds that

\[
\sum_{\tau=1}^{t} a_{\tau} \left( \langle y^{(\tau-1)} - y^{(\tau)}, x^* - x^{(\tau)} \rangle + h(y^{(\tau)}) - h(x^*) \right) - d(x^*) \]

\[ \leq \frac{\mu}{2} \sum_{\tau=1}^{t} a_{\tau} \left( \|x^{(\tau-1)} - x^*\|^2 - \|y^{(\tau-1)} - y^{(\tau)}\|^2 \right) + \frac{1}{2} \sum_{\tau=1}^{t} (1 + \mu A_{\tau-1}) \|y^{(\tau)} - y^{(\tau-1)}\|^2. \]  
(37)

**Proof of Lemma 3:** Due to limited page, the proof is omitted here and can be found in [32, Appendix D]. \( \blacksquare \)

Equipped with the aforementioned two technical lemmas, we are ready to present the proof of Theorem 2.

**Proof of Theorem 2:** For all \( \tau \geq 0 \), one has

\[
\frac{1}{n} \sum_{i=1}^{n} a_{\tau} \left( f_i(y^{(\tau)}) - f_i(x^*) \right) \leq \frac{1}{n} \sum_{i=1}^{n} a_{\tau} \left( f_i(x_i^{(\tau-1)}) - f_i(x^*) + \frac{L}{2} \|y^{(\tau)} - x_i^{(\tau-1)}\|^2 \right. \\
+ \left. \langle \nabla f_i(x_i^{(\tau-1)}), y^{(\tau)} - x_i^{(\tau-1)} \rangle \right) \]

\[ \leq \frac{1}{n} \sum_{i=1}^{n} a_{\tau} \left( \frac{L}{2} \|y^{(\tau)} - x_i^{(\tau-1)}\|^2 - \frac{\mu}{2} \|x_i^{(\tau-1)} - x^*\|^2 \right. \\
+ \left. \langle \nabla f_i(x_i^{(\tau-1)}), y^{(\tau)} - x_i^{(\tau-1)} \rangle \right) \]

\[ = \frac{1}{n} \sum_{i=1}^{n} a_{\tau} \left( \frac{L}{2} \|y^{(\tau)} - x_i^{(\tau-1)}\|^2 - \frac{\mu}{2} \|x_i^{(\tau-1)} - x^*\|^2 \right. \\
+ \left. \langle \nabla f_i(x_i^{(\tau-1)}), y^{(\tau)} - x^* \rangle \right) \]

\[ + a_{\tau} \left( \langle \overline{y}^{(\tau-1)}, y^{(\tau)} - x^* \rangle \right) \]  
(38)

where the two inequalities follow from (35) and (3), respectively, and the equality uses the definition of \( \overline{y}^{(\tau-1)} \). Upon summing up (38) from \( \tau = 0 \) to \( \tau = t \) and using Lemma 3 and \( F = \frac{1}{n} \sum_{i=1}^{n} f_i + h \), we obtain

\[
\sum_{\tau=1}^{t} a_{\tau} \left( F(y^{(\tau)}) - F(x^*) \right) \]

\[ \leq \frac{1}{n} \sum_{\tau=1}^{t} \sum_{i=1}^{n} a_{\tau} \left( \frac{L}{2} \|y^{(\tau)} - x_i^{(\tau-1)}\|^2 - \frac{\mu}{2} \|x_i^{(\tau-1)} - x^*\|^2 \right) \]

\[ + \frac{\mu}{2} \sum_{\tau=1}^{t} \sum_{i=1}^{n} a_{\tau} \left( \|x_i^{(\tau-1)} - x^*\|^2 - \|\overline{y}^{(\tau-1)} - y^{(\tau)}\|^2 \right) \]

\[ - \frac{1}{2} \sum_{\tau=1}^{t} (1 + \mu A_{\tau-1}) \|y^{(\tau)} - y^{(\tau-1)}\|^2 + d(x^*). \]  
(39)

Using the definition of \( \Delta y^{(\tau-1)} \) and the fact

\[ \|x^{(\tau-1)} - x^*\|^2 = \left\| \frac{1}{n} \sum_{i=1}^{n} x_i^{(\tau-1)} - x^* \right\|^2 \]

\[ \leq \frac{1}{n} \sum_{i=1}^{n} \|x_i^{(\tau-1)} - x^*\|^2 \]
one can simplify the aforementioned inequality to
\[
\sum_{\tau=1}^{t} a_{\tau} \left( F\left(y^{(\tau)} \right) - F(x^*) \right) \\
\leq \frac{1}{n} \sum_{\tau=1}^{t} \sum_{i=1}^{n} a_{\tau} \left( \frac{L}{2} \|y^{(\tau)} - x^{(\tau-1)}\|_2^2 \right) \\
- \frac{1}{n} \sum_{\tau=1}^{t} \frac{1 + \mu A_{\tau-1}}{2} \|\Delta y^{(\tau-1)}\|_2^2 + d(x^*).
\]
(40)

By the definition of \(\pi^{(\tau-1)}, x^{(\tau)},\) and \(y^{(\tau)},\) one can verify
\[
\sum_{i=1}^{n} \|y^{(\tau)} - \pi^{(\tau-1)}\|_2^2 \\
= \sum_{i=1}^{n} \left( \|\pi^{(\tau-1)}\|_2^2 - \|x_i^{(\tau-1)}\|_2^2 \right) \\
+ \sum_{i=1}^{n} \|y^{(\tau)} - x_i^{(\tau-1)}\|_2^2 \\
\geq \sum_{i=1}^{n} \left( \|\pi^{(\tau-1)} - y^{(\tau-1)}\|_2^2 - \|x_i^{(\tau-1)} - y^{(\tau-1)}\|_2^2 \right) \\
= \|y^{(\tau)} - x^{(\tau-1)}\|_2^2 - \|y^{(\tau-1)} - x^{(\tau-1)}\|_2^2.
\]

Besides, recall that \(F\) is convex, \(\tilde{y}^{(t)} = A_t^{-1} \sum_{\tau=1}^{t} a_{\tau} y^{(\tau)},\) and \(A_t = \sum_{\tau=1}^{t} a_{\tau}.\) These, together with (40), yield
\[
A_t \left( F(\tilde{y}^{(t)}) - F(x^*) \right) \\
\leq \frac{1}{n} \sum_{\tau=1}^{t} a_{\tau} \left( \frac{L}{2} \|y^{(\tau)} - x^{(\tau-1)}\|_2^2 \right) \\
- \frac{1}{n} \sum_{\tau=1}^{t} \frac{1 + \mu A_{\tau-1}}{2} \|\Delta y^{(\tau-1)}\|_2^2 + d(x^*).
\]

Upon using the inequality
\[
\|y^{(\tau)} - x^{(\tau-1)}\|_2^2 \leq 2 \|\Delta y^{(\tau-1)}\|_2^2 + 2 \|y^{(\tau-1)} - x^{(\tau-1)}\|_2^2
\]
we further obtain
\[
A_t \left( F(\tilde{y}^{(t)}) - F(x^*) \right) \\
\leq \frac{1}{n} \sum_{\tau=1}^{t} a_{\tau} \left( \frac{L}{2} - \frac{1 + \mu A_{\tau-1}}{2a_{\tau}} \right) \|\Delta y^{(\tau-1)}\|_2^2 \\
+ \frac{2L - \mu}{2n} \|x^{(\tau-1)} - y^{(\tau-1)}\|_2^2 + d(x^*) \\
= \frac{2L - \mu}{2n} \sum_{\tau=1}^{t} a_{\tau} \|\Delta y^{(\tau-1)}\|_2^2
\]
where the equality follows from the identity \(\frac{1 + \mu A_{\tau-1}}{2a_{\tau}} = \frac{1 - a_{\tau} \mu}{a_{\tau}},\)
which holds due to the update rule of \(\{a_{\tau}\}_{\tau \geq 0}\) and \(\{A_{\tau}\}_{\tau \geq 0}.\)
Upon taking expectation on both sides of the aforementioned inequality and using Lemma 2, one has
\[
A_t \left( \mathbb{E}[F(\tilde{y}^{(t)})] - F(x^*) \right) + \frac{\gamma}{2} \sum_{\tau=1}^{t} a_{\tau} \mathbb{E}[\|\Delta y^{(\tau-1)}\|_2^2] \\
\leq d(x^*) + \frac{(2L - \mu) \sigma^2}{n \theta (L + \mu)^2} = C
\]
(41)

where \(\gamma > 0\) is defined in (18). This implies (19) as desired. Moreover, it follows from (41) and \(A_t(\mathbb{E}[F(\tilde{y}^{(t)})] - F(x^*)) \geq 0\) that
\[
\sum_{\tau=1}^{t} a_{\tau} \mathbb{E}[\|\Delta y^{(\tau-1)}\|_2^2] \leq \frac{2nC}{\gamma}.
\]
This, together with the convexity of \(\cdot \|\cdot\|_2^2,\) Jensen’s Inequality, \(a_t \leq a_{t+1}\) for all \(t \geq 0,\) and Lemma 2, yields
\[
A_t \mathbb{E}[\|\tilde{x}^{(t)} - \bar{y}^{(t)}\|_2^2] \leq \sum_{\tau=1}^{t} a_{\tau} \mathbb{E}[\|x^{(\tau)} - y^{(\tau)}\|_2^2] \\
\leq \sum_{\tau=0}^{t} a_{\tau+1} \mathbb{E}[\|x^{(\tau)} - y^{(\tau)}\|_2^2] \\
\leq \frac{2}{\eta} \sum_{\tau=0}^{t-1} a_{\tau+1} \mathbb{E}[\|\Delta y^{(\tau)}\|_2^2] + \frac{2a \sigma^2}{n \theta (L + \mu)^2} \\
\leq \frac{4nC}{\eta \gamma} + \frac{2a \sigma^2}{\theta (L + \mu)^2} = D
\]
which implies (20) as desired.

\section*{Appendix C}
\section*{Proofs of Corollaries 1 and 2}
In this section, we provide the proofs of Corollaries 1 and 2.

\subsection*{A. Proof of Corollary 1}
\textit{Proof of Corollary 1}: Since \(\mu > 0,\) we obtain from the update of \(A_t\) in Algorithm 1 that
\[
\frac{1}{A_t} = \left( \frac{\mu}{1 - a_{\mu}} \right) \frac{1 - (1 - a_{\mu})^t}{a_{\mu}}.
\]

Besides, upon using the fact that \(F\) is strongly convex with modulus \(\mu,\) one has that for all \(t \geq 0\) and \(i = 1, \ldots, n\)
\[
\|x_i^{(t)} - x^*\|_2^2 \leq 2 \|x_i^{(t)} - y^{(t)}\|_2^2 + 2 \|y^{(t)} - x^*\|_2^2 \\
\leq 2 \|x_i^{(t)} - y^{(t)}\|_2^2 + \frac{1}{\mu} (F(y^{(t)}) - F(x^*)).
\]
These, together with (19) and (20), yields (21).
B. Proof of Corollary 2

Proof of Corollary 2: The upper bounds in (22) and (23) directly follow from the results in Theorem 2 and

$$\frac{1}{A_t} = \frac{1}{at}.$$  

For the special case $h(x) = 0$ and $d(x) = \frac{1}{2}||x||^2$, we consider

$$F(x^{(\tau)}) - F(y^{(\tau)}) = f(x^{(\tau)}) - f(y^{(\tau)})$$

$$\leq \frac{1}{n} \sum_{j=1}^{n} \left( f_j(x^{(\tau)}) - \langle \nabla f_j(x^{(\tau)}), y^{(\tau)} - x^{(\tau)} \rangle - f_j(x^{(\tau)}) \right)$$

$$\leq \frac{1}{n} \sum_{j=1}^{n} \left( \langle \nabla f_j(x^{(\tau)}), y^{(\tau)} - x^{(\tau)} \rangle - \langle \nabla f_j(x^{(\tau)}), y^{(\tau)} - x^{(\tau)} \rangle \right)$$

$$+ L \|y^{(\tau)} - y^{(\tau)}\|^2$$

$$+ L \|y^{(\tau)} - y^{(\tau)}\|^2$$

$$= \frac{1}{n} \sum_{j=1}^{n} \left( \langle \nabla f_j(x^{(\tau)}), x^{(\tau)} - y^{(\tau)} \rangle + L \|x^{(\tau)} - y^{(\tau)}\|^2 \right)$$

where the two inequalities follow from (3) and (35), respectively. The closed-form solutions for (12) and (17) can be derived as

$$x^{(\tau)} = -\frac{z^{(\tau)}}{1 + \mu A^t}, \quad y^{(\tau)} = -\frac{\tau^{(\tau)}}{1 + \mu A^t}.$$  

Therefore $y^{(\tau)} = \tau^{(\tau)}$. We sum up (42) from $i = 1$ to $i = n$ to get

$$\sum_{i=1}^{n} \left( F(x^{(\tau)}_i) - F(y^{(\tau)}_i) \right) \leq 2L \|y^{(\tau)} - x^{(\tau)}\|^2.$$  

Upon summing up (43) from $\tau = 1$ to $\tau = t$ and using the convexity of $F$, we obtain

$$t \sum_{i=1}^{n} \left( F(\tilde{x}^{(t)}_i) - F(y^{(t)}_i) \right) \leq \sum_{\tau=1}^{t} \sum_{i=1}^{n} \left( F(x^{(\tau)}_i) - F(y^{(\tau)}_i) \right)$$

$$\leq 2L \sum_{\tau=1}^{t} \|y^{(\tau)} - x^{(\tau)}\|^2$$

where $\tilde{x}^{(t)}_i = \frac{1}{t} \sum_{\tau=1}^{t} x^{(\tau)}_i$. After taking expectation on both sides of the aforementioned inequality and using Lemma 2 with $a_t = a$, we get

$$t \sum_{i=1}^{n} \mathbb{E} \left[ F(\tilde{x}^{(t)}_i) - F(y^{(t)}_i) \right]$$

$$\leq 4L \sum_{\tau=1}^{t} \mathbb{E}[\|\Delta y^{(\tau-1)}\|^2] + \frac{4L \sigma^2}{\theta(L + \mu)^2}$$

$$= 4L \sum_{\tau=1}^{t} \mathbb{E}[\|\Delta y^{(\tau-1)}\|^2] + \frac{4\sigma^2}{L(1 - \nu^2)}.$$  

By setting $\mu = 0$ and $d(x^*) = \frac{1}{2}||x^*||^2$ in (41), we have

$$\mathbb{E}[F(\tilde{y}^{(t)}_i)] - F(x^*)$$

$$\leq - \left( \frac{1}{2a} - L - \frac{2L}{(1 - \nu^2)} \right) \mathbb{E}[\|\Delta y^{(\tau-1)}\|^2]$$

$$+ \frac{n L \sigma^2}{2a \nu^2}.$$  

Also, by multiplying $n/a > 0$ on both sides of the aforementioned inequality and adding the resultant inequality to (45), we obtain

$$t \mathbb{E}[F(\tilde{x}^{(t)}_i)] - F(x^*) \leq t \sum_{i=1}^{n} \left( \mathbb{E}[F(\tilde{x}^{(t)}_i)] - F(x^*) \right)$$

$$\leq - \left( \frac{1}{2a} - L - \frac{6L}{(1 - \nu^2)} \right) \mathbb{E}[\|\Delta y^{(\tau-1)}\|^2]$$

$$+ \frac{n L \sigma^2}{2a \nu^2}.$$  

Now, using the condition in (24), we arrive at (25).

APPENDIX D

PROOFS OF SUPPORTING LEMMAS FOR THEOREM 2

A. Proof of Lemma 1

Proof of Lemma 1: We first show that $\nu$ monotonically increases with $a$ if $a \in (0, 1/\mu)$. Recall that $M$ is defined in (14). Then, the characteristic polynomial of $M$, denoted by $p(\lambda)$, is a quadratic function

$$p(\lambda) := \det(\lambda I - M) = (\lambda - M_{11})(\lambda - M_{22}) - M_{12}M_{21}$$

$$= \lambda^2 - \beta(2 + aL) \lambda - \frac{\beta^2}{1 - a \mu}.$$  

Using this, we obtain that $M$ has two real eigenvalues $\lambda_1 = (\xi_1 + \xi_2)/2$ and $\lambda_2 = (\xi_1 - \xi_2)/2$, where

$$\xi_1 = \frac{\beta(2 + aL)}{1 - a \mu}, \quad \xi_2 = \frac{\sqrt{a^2 \beta^2 L^2 + 4a \beta(\beta + 1)(L + \mu)}}{1 - a \mu}.$$  

Notice that $\xi_1 > 0$ and $\xi_2 > 0$ for any $a \in (0, 1/\mu)$. Thus, we have $\lambda_1 > 0$ and $|\lambda_1| > |\lambda_2|$ for any $a \in (0, 1/\mu)$. It then follows
that $\rho(M) = \lambda_1$ and

$$\nu(a) = \rho(M)\sqrt{1-a\mu} = \lambda_1\sqrt{1-a\mu}$$

$$= \frac{\beta(2+aL)}{2\sqrt{1-a\mu}} + \frac{\sqrt{\alpha^2 \beta^2 L^2 + 4\alpha \beta(\beta+1)(L+\mu)}}{2\sqrt{1-a\mu}}.$$  \hfill (50)

By routine calculation, one can verify that $\nu'(a) > 0$ if $a \in (0, \mu^{-1})$. Therefore, the value of $\nu$ monotonically increases with $a$ if $a \in (0, \mu^{-1})$.

Next, we show that $\nu < 1$ if (16) is satisfied. Note that (16) implies that $0 < a < \mu^{-1}$ and $\beta(2+3\mu)/(1-\beta)^2 < 1/a - \mu = (1-a\mu)/a$. It then follows that $1-a\mu \in (0, 1]$ and hence,

$$0 < (1-\beta)^2 - \frac{a\beta(2L+3\mu)}{1-a\mu} = 1 + \beta^2 - \left(\frac{\beta + a\beta(L+\mu)}{1-a\mu}\right) - \frac{a\beta(L+\mu)}{1-a\mu}.$$  

Upon dividing both sides of the aforementioned inequality by $1-a\mu$, we obtain

$$0 < \frac{1}{1-a\mu} + \frac{\beta^2}{1-a\mu} - \frac{a\beta(L+\mu)}{(1-a\mu)^2} \leq \frac{1}{1-a\mu} + \frac{\beta^2}{1-a\mu} - \frac{1}{\sqrt{1-a\mu}} \left(\frac{\beta + a\beta(L+\mu)}{1-a\mu}\right) - \frac{a\beta(L+\mu)}{(1-a\mu)^2}$$

$$= p' \left(1 + \frac{1}{\sqrt{1-a\mu}}\right)$$  \hfill (51)

where the second inequality is due to $1-a\mu \in (0, 1]$ and the equality follows from (48). Besides, using the definition of characteristic polynomial, one further has

$$0 < p' \left(1 + \frac{1}{\sqrt{1-a\mu}}\right) = \left(1 + \frac{1}{\sqrt{1-a\mu}} - M_{11}\right) \left(1 + \frac{1}{\sqrt{1-a\mu}} - M_{22}\right) - M_{12}M_{21}. \hfill (52)$$

By (14), $\beta \in (0, 1)$, and $1-a\mu \in (0, 1]$, we have that $M_{12} > 0$, $M_{21} > 0$, and $1/\sqrt{1-a\mu} > M_{11}$. It then follows that $1/\sqrt{1-a\mu} > M_{22}$, and hence,

$$\nu' \left(1 + \frac{1}{\sqrt{1-a\mu}}\right) = \frac{2}{1-a\mu} - M_{11} - M_{22} > 0.$$  

This, together with the fact that $q$ is a quadratic function, implies that $q(\lambda)$ is monotonically increasing on $[1/\sqrt{1-a\mu}, \infty)$. It then follows from (51) that $1/\sqrt{1-a\mu} > \lambda_1 = \rho(M)$, which implies that $\nu < 1$.

\subsection{Proof of Lemma 2}

In this subsection, we first present three technical lemmas, and then, provide the proof of Lemma 2.

\textbf{Lemma 4}: For the sequences $\{\tilde{x}^{(t)}\}_{t \geq 0}$ and $\{\tilde{z}^{(t)}\}_{t \geq 0}$ defined in (28), one has that for any $t \geq 0$,

$$\tilde{x}^{(t)} = \tilde{g}^{(t)} - \mu \tilde{x}^{(t)}, \quad \tilde{z}^{(t)} = \sum_{\tau=0}^{t-1} a_{t+1} \tilde{z}^{(\tau)}. \hfill (53)$$

\textbf{Proof}: We prove by an induction argument. Since $s_i^{(0)} = \nabla f_i(x^{(0)}) - \mu x^{(0)}, z_i^{(0)} = 0$, and $x_i^{(0)} = x^{(0)}$ for all $i$, we readily have that (53) holds when $t = 0$. Now, suppose that (53) holds for $t - 1$. From (27) and (28), we observe that the following identities hold for any $\tau \geq 0$:

$$\tilde{x}^{(\tau)} = \frac{1}{n}(1^T \otimes I) \tilde{x}^{(\tau)}, \quad \tilde{g}^{(\tau)} = \frac{1}{n}(1^T \otimes I) \nabla^{(\tau)}$$

$$\tilde{z}^{(\tau)} = \frac{1}{n}(1^T \otimes I) \tilde{z}^{(\tau)}, \quad \tilde{z}^{(\tau)} = \frac{1}{n}(1^T \otimes I) \tilde{z}^{(\tau)}.$$  \hfill (54)

It then follows from this and (32b) that

$$\tilde{x}^{(t)} = \frac{1}{n}(1^T \otimes I) s^{(t)}$$

$$= \frac{1}{n}(1^T \otimes I) (P^{(t-1)} \otimes I) s^{(t-1)} + \frac{1}{n}(1^T \otimes I) \nabla^{(t-1)}$$

$$= \frac{1}{n}(1^T \otimes I) s^{(t-1)} + \frac{1}{n}(1^T \otimes I) \nabla^{(t-1)}$$

$$\leq \frac{1}{n}(1^T \otimes I) s^{(t-1)} + \frac{1}{n}(1^T \otimes I) \nabla^{(t-1)} - \mu \tilde{x}^{(t-1)} + \mu \tilde{x}^{(t-1)}$$

$$= \tilde{x}^{(t-1)} + \tilde{z}^{(t-1)} - \mu \tilde{x}^{(t-1)} + \mu \tilde{x}^{(t-1)}$$

$$= \tilde{g}^{(t)} - \mu \tilde{x}^{(t)}$$

where the second equality is due to (32b), the third equality uses the fact that $(A \otimes B)(C \otimes D) = (AC \otimes BD)$, the fourth equality follows from the fact that $P^{(t-1)}$ is doubly stochastic, and the last equality is due to the assumption that (53) holds for $t - 1$. Similarly, by (32a) and (54), we obtain

$$\tilde{z}^{(t)} = \frac{1}{n}(1^T \otimes I) z^{(t)}$$

$$= \frac{1}{n}(1^T \otimes I) (P^{(t-1)} \otimes I) (z^{(t-1)} + a_t s^{(t-1)})$$

$$= \frac{1}{n}(1^T \otimes I) (z^{(t-1)} + a_t s^{(t-1)})$$

$$= \tilde{z}^{(t-1)} + a_t \tilde{s}^{(t-1)} = \sum_{\tau=0}^{t-1} a_{t+1} \tilde{z}^{(\tau)}.$$  \hfill (55)

Therefore, (53) holds for $t$ and the induction argument is completed.

\textbf{Lemma 5}: For the sequence $\{x_i^{(t)}\}_{i=1, \ldots, n, t \geq 0}$ generated by Algorithm 1 and the auxiliary sequence $\{y_i^{(t)}\}_{t \geq 0}$ defined in (17), one has that for all $t \geq 0$ and $i = 1, \ldots, n$,

$$\|x_i^{(t)} - y_i^{(t)}\| \leq \frac{1}{1 + \mu A_t} \|z_i^{(t)} - \tilde{z}^{(t)}\|$$  \hfill (55)
Proof: It is easy to see that (55) holds when \( t = 0 \) because both sides of (55) equal 0. Now, suppose that \( t \geq 1 \). Recall that \( d \) is strongly convex with modulus 1. Let the mapping \( R : \mathbb{R}^m \to \mathbb{R}^m \) be defined as

\[
R(\omega) := \arg\min_{x \in \mathbb{R}^m} \{(\omega, x) + \phi(x)\}
\]

where \( \phi(x) = A_1(x_1 + \mu_2 \|x\|^2/2 + h(x)) + d(x) \) is strongly convex with modulus 1 + \( \mu_2 A_1 \). Then, by (12) and (17), we have

\[
y^{(t)} = R(z^{(t)}), \quad x^{(t)} = R(z^{(t)}), \quad \forall i = 1, \ldots, n.
\]

Moreover, the mapping \( R \) is Lipschitz continuous with Lipschitz constant \( (1 + \mu_2 A_1)^{-1} \); see, e.g., [43, Prop. 4.9]. This immediately implies (55) as desired.

Next, we recall a lemma from [25, Lemma 4].

**Lemma 6:** Suppose that \( \{q^{(t)}\}_{t \geq 0} \) and \( \{p^{(t)}\}_{t \geq 0} \) are two sequences of positive scalars such that for all \( t \geq 0 \)

\[
q^{(t)} \leq \nu^t c + \sum_{\tau=0}^{t-1} \nu^{t-1-\tau} p^{(\tau)}
\]

where \( \nu \in (0, 1) \) and \( c \geq 0 \) is a constant. Then, the following holds for all \( t \geq 0 \):

\[
\sum_{\tau=1}^{t} (q^{(\tau)})^2 \leq \frac{2}{(1-\nu^2)} \sum_{\tau=0}^{t-1} (p^{(\tau)})^2 + \frac{2c^2}{1-\nu^2}.
\]

Now, we are ready to prove Lemma 2.

**Proof of Lemma 2:** From Lemma 4, we have

\[
\tilde{z}^{(t)} = \tilde{z}^{(t-1)} + a_t \tilde{s}^{(t-1)}.
\]

This, together with (32a) and the definition of \( \tilde{z}^{(t-1)} \) in (29), yields

\[
\tilde{z}^{(t)} = P^{(t-1)} \tilde{z}^{(t-1)} - (1 \otimes \tilde{s}^{(t-1)}) + a_t \left( P^{(t-1)} \tilde{s}^{(t-1)} - 1 \otimes \tilde{s}^{(t-1)} \right).
\]

(56)

It then follows from (34) that

\[
\begin{align*}
\|\tilde{z}^{(t)}\|_E & \leq \left\| P^{(t-1)} \tilde{z}^{(t-1)} - 1 \otimes \tilde{s}^{(t-1)} \right\|_E \\
& \quad + a_t \left\| P^{(t-1)} \tilde{s}^{(t-1)} - 1 \otimes \tilde{s}^{(t-1)} \right\|_E.
\end{align*}
\]

(57)

Note that \( 1 \otimes \tilde{z}^{(t-1)} = (1 \otimes I) \tilde{z}^{(t-1)} \), which, together with (54) and the identity \( (A \otimes B)(C \otimes D) = (AC \otimes BD) \), yields

\[
1 \otimes \tilde{z}^{(t-1)} = \frac{1}{n}(1 \otimes I)(1 \otimes I) \tilde{z}^{(t-1)} = \left( \frac{11^T}{n} \otimes I \right) \tilde{z}^{(t-1)}.
\]

Using this and \( P^{(t-1)} = P^{(t-1)} \otimes I \), we obtain

\[
\begin{align*}
P^{(t-1)} \tilde{z}^{(t-1)} & - 1 \otimes \tilde{z}^{(t-1)} \\
& = (P^{(t-1)} \otimes I) \tilde{z}^{(t-1)} - \left( \frac{11^T}{n} \otimes I \right) \tilde{z}^{(t-1)} \\
& = \left( P^{(t-1)} - \frac{11^T}{n} \right) \otimes I \tilde{z}^{(t-1)} \\
& = \left( \left( P^{(t-1)} - \frac{11^T}{n} \right) \otimes I \right) \tilde{z}^{(t-1)}
\end{align*}
\]

and

\[
\begin{align*}
P^{(t-1)} \tilde{z}^{(t-1)} & - 1 \otimes \tilde{z}^{(t-1)} \\
& = \left( P^{(t-1)} - \frac{11^T}{n} \right) \otimes I \tilde{z}^{(t-1)}
\end{align*}
\]

\[
= \left( \left( P^{(t-1)} - \frac{11^T}{n} \right) \otimes I \right) \tilde{z}^{(t-1)}
\]

where the third equality uses (29) and \( 1 \otimes \tilde{z}^{(t-1)} = (1 \otimes I) \tilde{z}^{(t-1)} \), the fourth equality follows from the identity \( (A \otimes B)(C \otimes D) = (AC \otimes BD) \), and the last one is due to the fact that \( P^{(t-1)} \) is doubly stochastic. Then, by (33) and Assumption 2, one has

\[
\begin{align*}
\left\| \left( P^{(t-1)} - \frac{11^T}{n} \right) \tilde{z}^{(t-1)} \right\|_E & \leq \beta \left\| \tilde{z}^{(t-1)} \right\|_E \\
& \leq \beta \left\| \tilde{s}^{(t-1)} \right\|_E.
\end{align*}
\]

(58)

It then follows from (57) that

\[
\left\| \tilde{z}^{(t)} \right\|_E \leq \beta \left( \left\| \tilde{z}^{(t-1)} \right\|_E + a_t \left\| \tilde{s}^{(t-1)} \right\|_E \right).
\]

(59)

Similarly, from Lemma 4, (29), and (32b), we obtain

\[
\begin{align*}
\left\| \tilde{s}^{(t)} \right\|_E & \leq \left\| P^{(t-1)} \tilde{s}^{(t-1)} - (1 \otimes I) \tilde{s}^{(t-1)} \right\|_E \\
& \quad + \left\| \nabla^{(t)} - \nabla^{(t-1)} - (1 \otimes I) (\tilde{g}^{(t)} - \tilde{g}^{(t-1)}) \right\|_E \\
& \quad + \mu \left\| (1 \otimes I) \Delta \tilde{z}^{(t-1)} - \Delta x^{(t-1)} \right\|_E.
\end{align*}
\]

(60)

By (54), one can verify that

\[
\begin{align*}
\nabla^{(t)} - \nabla^{(t-1)} - (1 \otimes I) (\tilde{g}^{(t)} - \tilde{g}^{(t-1)}) & = \left( \left( I - \frac{11^T}{n} \right) \otimes I \right) \left( \nabla^{(t)} - \nabla^{(t-1)} \right),
\end{align*}
\]

\[
\Delta x^{(t-1)} - (1 \otimes I) \Delta \tilde{z}^{(t-1)} = \left( \left( I - \frac{11^T}{n} \right) \otimes I \right) \Delta x^{(t-1)}
\]

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
which, respectively, imply that
\[
\|\nabla^{(r)} - \nabla^{(r-1)} - (1 \otimes I)(\tilde{g}^{(r)} - \tilde{g}^{(r-1)})\|_E \\
\leq \left\| \left( I - \frac{1}{n} I \right) \otimes I \right\| \|\nabla^{(r)} - \nabla^{(r-1)}\|_E \leq \|\nabla^{(r)} - \nabla^{(r-1)}\|_E,
\]
\[
\|\Delta x^{(k-1)} - (1 \otimes I) \Delta x^{(k-1)}\|_E \\
\leq \left\| \left( I - \frac{1}{n} I \right) \otimes I \right\| \|\Delta x^{(k-1)}\|_E \leq \|\Delta x^{(k-1)}\|_E.
\]
Besides, it follows from (4) that \(\|\nabla^{(r)} - \nabla^{(r-1)}\|_E \leq L\|\Delta x^{(k-1)}\|_E\). Upon substituting these and (58) into (60), we obtain
\[
\|\tilde{s}^{(r)}\|_E \leq \beta \|\tilde{s}^{(r-1)}\|_E + (L + \mu)\|\Delta x^{(r-1)}\|_E.
\]  
(61)

Multiplying the both sides of the aforementioned inequality by \(a_{\tau+1} = a_{\tau} / (1 - a_{\mu})\), we have
\[
a_{\tau+1}\|\tilde{s}^{(r)}\|_E \\
\leq \frac{1}{1 - a_{\mu}} \left( a_{\tau} \|\tilde{s}^{(r-1)}\|_E + (L + \mu)a_{\tau}\|\Delta x^{(r-1)}\|_E \right).
\]  
(62)

Upon using Lemma 5 and
\[
\|\Delta x^{(r-1)}\|_E \\
\leq \|x^{(r)} - y^{(r)}\|_E + \|x^{(r-1)} - y^{(r-1)}\|_E + \|\Delta y^{(r-1)}\|_E
\]
on one has
\[
a_{\tau}\|\Delta x^{(r-1)}\|_E \\
\leq \frac{a_{\tau}}{1 + a_{\mu} A_{\tau}} \|\tilde{z}^{(r)}\|_E + \frac{a_{\tau}}{1 + a_{\mu} A_{\tau}} \|\tilde{z}^{(r-1)}\|_E + a_{\tau}\|\Delta y^{(r-1)}\|_E
\]
where the equality follows from
\[
\frac{1 + a_{\mu} A_{\tau}}{a_{\tau}} = \frac{\left( \frac{1}{1 - a_{\mu}} \right)^{\tau} - 1}{a}
\]
In light of (59), we have
\[
a_{\tau}\|\Delta x^{(r-1)}\|_E \leq \left( a_{\beta} + \frac{a_{\tau}}{1 - a_{\mu}} \right) \|\tilde{z}^{(r-1)}\|_E \\
+ a_{\beta} a_{\tau}\|\tilde{s}^{(r-1)}\|_E + a_{\tau}\|\Delta y^{(r-1)}\|_E.
\]
Therefore,
\[
a_{\tau+1}\|\tilde{s}^{(r)}\|_E \leq \frac{\beta + a_{\beta}(L + \mu)}{1 - a_{\mu}} a_{\tau}\|\tilde{s}^{(r-1)}\|_E \\
+ \frac{a(L + \mu)}{1 - a_{\mu}} \left( \beta + \frac{1}{1 - a_{\mu}} \right) \|\tilde{z}^{(r-1)}\|_E \\
+ \frac{L + \mu}{1 - a_{\mu}} a_{\tau}\|\Delta y^{(r-1)}\|_E.
\]  
(63)

By combining (59) and (63), the following inequality can be established:
\[
\begin{bmatrix}
\|\tilde{z}^{(r)}\|_E \\
\|\tilde{s}^{(r)}\|_E
\end{bmatrix}
\leq M \begin{bmatrix}
\|\tilde{z}^{(r-1)}\|_E \\
\|\tilde{s}^{(r-1)}\|_E
\end{bmatrix}
+ L + \mu \begin{bmatrix}
\frac{1}{1 - a_{\mu}} a_{\tau}\|\Delta y^{(r-1)}\|_E
\end{bmatrix}
\]
where \(M\) is defined in (14). By iterating the preceding linear system inequality and using
\[
\begin{bmatrix}
\|\tilde{z}^{(0)}\|_E \\
\|\tilde{s}^{(0)}\|_E
\end{bmatrix} = \sqrt{\sum_{i=1}^{n} \left( \|\tilde{w}_i\| - \bar{g}^{(0)} \right)^2} = \sigma
\]
we obtain
\[
\begin{bmatrix}
\|\tilde{z}^{(t)}\|_E \\
\|\tilde{s}^{(t)}\|_E
\end{bmatrix} \\
\leq M^t \begin{bmatrix}
0 \\
0
\end{bmatrix}
+ \begin{bmatrix}
\frac{L + \mu}{1 - a_{\mu}} \sum_{\tau=0}^{t-1} M^{t-\tau-1} \sqrt{a_{\tau+1}} \|\Delta y^{(\tau)}\|_E
\end{bmatrix}
\]
Recall that the eigenvalues for the matrix \(M\) are \(\lambda_1 = (\xi_1 + \xi_2)/2\) and \(\lambda_2 = (\xi_1 - \xi_2)/2\), where
\[
\xi_1 = \beta(2 + aL), \quad \xi_2 = \frac{a(L + \mu)}{1 - a_{\mu}} \sqrt{\frac{\beta^2 L^2}{(L + \mu)^2} + \frac{4\beta(\beta + 1)}{a(L + \mu)}}.
\]
Thus, the analytical form for the \(n\)th power of \(M\) is (see, e.g., [44])
\[
M^n = \lambda_1^n \left( \frac{M - \lambda_2 I}{\lambda_1 - \lambda_2} \right) + \lambda_2^n \left( \frac{M - \lambda_1 I}{\lambda_1 - \lambda_2} \right)
\]
It then follows that
\[
\begin{bmatrix}
M^n \end{bmatrix}_{12} = \frac{M_{12}(\lambda_1^n - \lambda_2^n)}{\lambda_1 - \lambda_2} = \frac{\beta(\lambda_1^n - \lambda_2^n)}{\lambda_1 - \lambda_2} \leq \frac{2\beta(r(M))^n}{\xi_2}
\]
where \(r(M)\) is the spectral radius of \(M\). Due to our assumption that
\[
1 > \frac{\beta(2L + 3\mu)}{(1 - \beta)^2} + \mu > \frac{2\beta(L + \mu)}{(1 - \beta)^2}
\]
and \(\beta \in (0, 1)\), we have \(\xi_2 > 2a\beta(L + \mu) / (1 - a_{\mu})\). Therefore,
\[
\|\tilde{z}^{(t)}\|_E \leq \frac{2\beta L + \mu}{1 - a_{\mu}} \sqrt{\sum_{\tau=0}^{t-1} \|\Delta y^{(\tau)}\|_E}
\]
This bound, together with Lemma 5, yields
\[
\sqrt{\alpha_{t+1}} \left\| y^{(t)} \right\|_E 
\leq \sqrt{\frac{\sqrt{\alpha_{t+1}}}{\mu}} \sum_{\tau=0}^{t-1} \sqrt{\frac{\mu}{\tau+1}} \left\| \Delta y^{(\tau)} \right\|_E + \left( \frac{\rho(M)^t}{L} \right) \sqrt{\frac{1}{\alpha_{t+1}}} \left( \frac{\rho(M)^t}{L} \mu \right) \right)
\]
\[
\leq \frac{1}{\sqrt{\alpha_{t+1}}} \left( \frac{\sqrt{\alpha_{t+1}}}{\mu} \sum_{\tau=0}^{t-1} \sqrt{\frac{\mu}{\tau+1}} \left\| \Delta y^{(\tau)} \right\|_E + \left( \frac{\rho(M)^t}{L} \right) \sqrt{\frac{1}{\alpha_{t+1}}} \left( \frac{\rho(M)^t}{L} \mu \right) \right).
\]
(65)

The desired inequality (36) then follows from this and Lemma 6.

REFERENCES

1. W. Shi, Q. Ling, G. Wu, and W. Yin, “A proximal gradient algorithm for decentralized composite optimization,” IEEE Trans. Signal Process., vol. 63, no. 22, pp. 6013–6023, Nov. 2015.
2. S. Alghunaim, K. Yuan, and A. H. Sayed, “A linearly convergent proximal gradient algorithm for decentralized optimization,” in Proc. Adv. Neural Inf. Process. Syst., 2019, pp. 2848–2858.
3. R. L. Raffard, C. J. Tomlin, and S. P. Boyd, “Distributed optimization for resource allocation,” in Proc. 49th Allerton Conf. Commun. Control. Comput., 2011, pp. 434–436.
4. X. Liu, C. Zhang, H. Zhang, C. J. Hsieh, W. Zhang, and J. Liu, “Can decentralized algorithms outperform centralized algorithms? A case study for decentralized parallel stochastic gradient descent,” in Proc. 17th Conf. Decis. Control, 2018, pp. 7330–7334.
5. T. Yang et al., “A survey of distributed optimization,” Annu. Rev. Control, vol. 47, pp. 278–305, 2019.
6. S. Boyd, A. Gosh, B. Prabhakar, and D. Shah, “Randomized gossip algorithms,” IEEE Trans. Inf. Theory, vol. 52, no. 6, pp. 2508–2530, Jun. 2006.
7. K. I. Tsianos, S. Lawlor, and M. G. Rabbat, “Push-sum distributed dual averaging for convex optimization,” in Proc. 51st IEEE Conf. Decis. Control, 2012, pp. 5453–5458.
8. S. Lee, A. Nedić, and M. Raginsky, “Coordinate dual averaging for decentralized online optimization with nonseparable global objectives,” IEEE Trans. Control Netw. Syst., vol. 5, no. 1, pp. 34–44, Mar. 2018.
9. I. Colin, A. Bellet, J. Salmon, and S. Clémençon, “Gossip dual averaging for decentralized optimization of pairwise functions,” in Proc. 33rd Int. Conf. Mach. Learn., 2016, pp. 1388–1396.
10. C. Liu, H. Li, and Y. Shi, “Towards an $O(1/t)$ convergence rate for distributed dual averaging,” IFAC-PapersOnLine, vol. 53, no. 2, pp. 3254–3259, 2020.
11. S. A. Alghunaim, E. K. Ryu, K. Yuan, and A. H. Sayed, “Decentralized proximal gradient algorithms with linear convergence rates,” IEEE Trans. Autom. Control, vol. 66, no. 6, pp. 2787–2794, Jun. 2021.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Changxin Liu (Member, IEEE) received the Ph.D. degree in mechanical engineering from the University of Victoria, Victoria, BC, Canada, in 2021. Since October 2021, he has been a Postdoctoral Researcher with the School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden. His research interests include distributed optimization and control of networked systems.

Dr. Liu is an active reviewer for more than ten international journals and conferences, and was an Outstanding Reviewer for IEEE TRANSACTIONS ON CYBERNETICS in 2018.

Yong Zhang received his Ph.D. degree from Simon Fraser University, Burnaby, BC, Canada, in 2014. He is currently a Distinguished Researcher with Huawei Technologies Canada, Burnaby, BC, Canada and leading the Big Data and Intelligence Platform Lab, Vancouver Research Center. Prior to that, he was a Postdoctoral Research Fellow with Stanford University, Stanford, CA, USA. His research works have been published in top-tier journals and conferences. His research interests include large-scale numerical optimization and machine learning.

Zirui Zhou received the Ph.D. degree in systems engineering and engineering management from the Chinese University of Hong Kong, Hong Kong, in 2015. He is currently a Senior Principal Researcher with Huawei Technologies Canada, Burnaby, BC, Canada. Before joining Huawei, he was an Assistant Professor with the Department of Mathematics at Hong Kong Baptist University. His research interests include optimization solvers, numerical optimization, and its applications in machine learning. His research works on convex analysis, optimization theory, and provable nonconvex methods have been published in top-tier journals and conferences.

Jian Pei (Fellow, IEEE) received the Ph.D. degree in computing science from Simon Fraser University, Burnaby, BC, Canada, in 2002. He is currently a Professor with the School of Computing Science and an Associate Member with the Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, Canada. His research interests include data science, Big Data, data mining, and database systems. His expertise is on developing effective and efficient data analysis techniques for novel data intensive applications.

Prof. Pei is recognized as a Fellow of the Royal Society of Canada (RSC), Academy of Science, and a Fellow of the ACM.

Yang Shi (Fellow, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Alberta, Edmonton, AB, Canada, in 2005. From 2005 to 2009, he was an Assistant Professor and an Associate Professor with the Department of Mechanical Engineering, University of Saskatchewan, Saskatoon, SK, Canada. In 2009, he joined the University of Victoria, Victoria, BC, Canada, where he is currently a Professor with the Department of Mechanical Engineering. He was a Visiting Professor with the University of Tokyo, Tokyo, Japan, in 2013. His current research interests include networked and distributed systems, model predictive control (MPC), cyber-physical systems (CPS), robotics and mechatronics, navigation and control of autonomous systems (AUV and UAV), and energy system applications.

Prof. Shi was the recipient of the University of Saskatchewan Student Union Teaching Excellence Award in 2007; the Faculty of Engineering Teaching Excellence Award in 2012 and the Craigdarroch Silver Medal for Excellence in Research in 2015 from the University of Victoria; the 2017 IEEE TRANSACTIONS ON FUZZY SYSTEMS Outstanding Paper Award for his coauthored paper; the JSBSF Invitation Fellowship (short-term); and the Humboldt Research Fellowship for Experienced Researchers in 2018. He has been a Member of the IEEE IES Administrative Committee since 2017 and is currently the Chair of IEEE IES Technical Committee on Industrial Cyber-Physical Systems. He is the Co-Editor-in-Chief of IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, and also serves as an Associate Editor for Automatica and IEEE TRANSACTIONS ON CYBERNETICS. He is a Fellow of ASME, Engineering Institute of Canada (EIC), and Canadian Society for Mechanical Engineering (CSME), and a Registered Professional Engineer in British Columbia, Canada.