Rate analysis of dual averaging for nonconvex distributed optimization

Changxin Liu * **Xuyang Wu** * **Xinlei Yi** * **Yang Shi** **
Karl H. Johansson *

* *School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, and Digital Futures, 100 44 Stockholm, Sweden (e-mail: {changxin; xuyangw; xinleiy; kallej}@kth.se).*
** *Department of Mechanical Engineering, University of Victoria, Victoria, B.C. V8W 3P6, Canada (e-mail: yshi@uvic.ca).*

Abstract: This work studies nonconvex distributed constrained optimization over stochastic communication networks. We revisit the distributed dual averaging algorithm, which is known to converge for convex problems. We start from the centralized case, for which the change of two consecutive updates is taken as the suboptimality measure. We validate the use of such a measure by showing that it is closely related to stationarity. This equips us with a handle to study the convergence of dual averaging in nonconvex optimization. We prove that the squared norm of this suboptimality measure converges at rate $\mathcal{O}(1/t)$. Then, for the distributed setup we show convergence to the stationary point at rate $\mathcal{O}(1/t)$. Finally, a numerical example is given to illustrate our theoretical results.

Keywords: Dual averaging, nonconvex optimization, distributed constrained optimization, stochastic networks, multi-agent consensus.

1. INTRODUCTION

In recent years, distributed optimization has received surged research interests from both academia and industry, because of its capability of delivering high-quality solutions to a system-wide task under the support of a cluster of computing units/agents and real-time communication networks. For a recent overview of distributed optimization, the interested readers are referred to (Yang et al., 2019).

This work is concerned with the distributed optimization problem where the cost function is the sum of multiple smooth and possibly nonconvex objective functions locally with the agents, the constraint set is common across the agents, and the communication network is time-varying and random. Such formulation finds wide applications including platooning control of multiple vehicles (Shen et al., 2022), machine learning (Lian et al., 2017), to name a few. Particularly, the stochastic time-varying communication network is of practical significance because real communication networks suffer from congestion, failure, and random package dropouts.

Existing works on distributed nonconvex optimization mostly dealt with fixed communication networks; see, e.g., (Di Lorenzo and Scutari, 2015; Hong et al., 2017; Yi et al., 2021). Recently, Scutari and Sun (2019); Xin et al. (2021); Jiang et al. (2022) considered nonconvex composite optimization with deterministic time-varying networks. However, the communication network is essentially assumed to be connected in every finite steps. Note that all the above methods are developed based on gradient descent. Different from them, distributed dual averaging (DDA) originally proposed by Duchi et al. (2011) has demonstrated its advantages in simultaneously handling constraints and stochastic communication networks. Nevertheless, this type of algorithms were only known to converge for convex problems to the best of our knowledge.

It is worth mentioning that there are a few recent attempts in the literature regarding the convergence of centralized dual averaging (CDA) for nonconvex optimization. For example, Defazio and Jelassi (2022) established the relation between hyperparameters in CDA and stochastic gradient descent (SGD), and then generalized the analysis in SGD to CDA. However, the analysis is limited to unconstrained optimization, in which SGD and CDA only differ in the choice of hyperparameters. However, they may generate distinct trajectories of variables in the presence of constraints (Fang et al., 2022). Héliou et al. (2020) investigated the behavior of dual averaging in online nonconvex optimization with constraints. The authors considered nonsmooth time-varying loss functions with bounded subgradients, which is not applicable to the setup considered in this work.

In this work, we extend the dual averaging based distributed optimization algorithm developed in (Liu et al., 2022a) to nonconvex constrained problems. The main contributions of this work are as follows. First, we prove the convergence rate of CDA for nonconvex smooth optimization with constraints for the first time. A new measure of suboptimality is defined and its relation to stationarity is discussed. Based on them, we prove the $\mathcal{O}(1/t)$ con-

⋆ This work was supported by the Knut and Alice Wallenberg Foundation, the Swedish Foundation for Strategic Research, and the Swedish Research Council.

vergence rate of dual averaging in terms of the squared norm of the suboptimality measure. Then, the results are extended to the distributed setup with stochastic communication networks. Under rather mild conditions, the convergence rate of DDA is proved to be $\mathcal{O}(1/t)$.

Notation Given a convex set $\mathcal{X} \subset \mathbb{R}^m$, we denote the normal cone to $\mathcal{X}$ at x by $\mathcal{N}_{\mathcal{X}}(x) = \{g \in \mathbb{R}^m : \langle g, y - x\rangle \leq 0, \forall y \in \mathcal{X}\}$. For a real-valued random vector x, we define $\|x\|_{\mathbb{E}} = \sqrt{\mathbb{E}\|x\|^2}$. We use $\rho(\cdot)$ to denote the spectral radius of a matrix. A differentiable function d is said strongly convex with modulus $a^{-1} > 0$ if

$$d(y) \geq d(x) + \langle \nabla d(x), y - x\rangle + \frac{1}{2a}\|y - x\|^2, \ \forall x, y.$$

2. PROBLEM STATEMENT

2.1 Optimization problem

Consider the finite-sum constrained optimization problem

$$\min_{x\in\mathcal{X}} \left\{ f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) \right\} \tag{1}$$

where each f_i is a smooth and possibly nonconvex function, and $\mathcal{X} \subset \mathbb{R}^m$ is a compact convex set. The optimal objective value is denoted as $f^* > -\infty$.

Assumption 1. Each f_i is continuously differentiable on an open set that contains $\mathcal{X}$, and ∇f_i is Lipschitz continuous with Lipschitz constant $L > 0$, i.e.,

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x - y\|, \quad \forall x, y \in \mathcal{X}.$$

A direct consequence of Assumption 1 is

$$f_i(y) \leq f_i(x) + \langle \nabla f_i(x), y - x\rangle + \frac{L}{2}\|y - x\|^2, \quad \forall x, y \in \mathcal{X}.$$

For Problem (1), we recall the stationarity condition, which is a necessary local optimality condition (Beck, 2017, Definition 3.73).

Definition 1. (stationary point). A point $x^* \in \mathcal{X}$ is a stationary point of Problem (1) if $-\nabla f(x^*) \in \mathcal{N}_{\mathcal{X}}(x^*)$.

2.2 Communication network

Consider the standard distributed optimization setup, where each agent i holds a local objective function f_i and is only able to communicate with other agents if they are connected in the communication network. At time t, a doubly stochastic matrix $P^{(t)}$ is used to describe the network topology and the weights of connected links. In this work, we consider a general setting of stochastic communication networks, i.e., $P^{(t)}$ is a random matrix for every t. We denote by $p_{ij}^{(t)}$ the (i,j)-th element in $P^{(t)}$. $p_{ij}^{(t)} > 0$ only if the two agents i and j are neighbors at t. The set of i's neighbors at time t is denoted as $\mathcal{N}_i^{(t)}$.

Assumption 2. For every $t \geq 0$, it holds that i) $P^{(t)}\mathbf{1} = \mathbf{1}$ and $\mathbf{1}^T P^{(t)} = \mathbf{1}^T$, where $\mathbf{1}$ denotes the all-one vector of dimensionality n; ii) $P^{(t)}$ is independent of the random events that occur up to time $t-1$; and iii) there exists a constant $\beta \in (0,1)$ such that

$$\sqrt{\rho\left(\mathbb{E}_t\left[{P^{(t)}}^T P^{(t)}\right] - \frac{\mathbf{1}\mathbf{1}^T}{n}\right)} \leq \beta, \tag{2}$$

where the expectation $\mathbb{E}_t[\cdot]$ is taken with respect to the distribution of $P^{(t)}$ at time t.

Assumption 2 is satisfied by a host of common stochastic networks, e.g., randomized gossip (Boyd et al., 2006) and Bernoulli stochastic networks (Kar and Moura, 2008). Different from the deterministic time-varying networks considered in (Nedic et al., 2017; Xin et al., 2021; Jiang et al., 2022), Assumption 2 does not require the communication network to be connected every finite time steps. In fact, for stochastic networks defined in Assumption 2, it is possible that the mixing matrix $P^{(t)}$ never produces a deterministic contraction property in finite steps. Thus, the convergence analysis for deterministic time-varying networks cannot be applied or easily extended to the setting of stochastic networks.

This work focuses on the theoretical convergence properties of dual averaging algorithms for nonconvex optimization in both centralized and distributed settings.

3. DUAL AVERAGING ALGORITHM FOR NONCONVEX OPTIMIZATION

In this section, we present the CDA algorithm and derive its convergence rates for nonconvex problems.

Given constant $a > 0$ and an arbitrary variable $x^{(0)} \in \mathcal{X}$, we define a class of proximal functions $d : \mathbb{R}^m \to \mathbb{R}$.

Definition 2. (proximal function). d is called a proximal function if: i) $x^{(0)}$ is the "$d(\cdot)$-center" of $\mathcal{X}$, i.e., $x^{(0)} = \operatorname{argmin}_{x\in\mathcal{X}} d(x)$ and $d(x^{(0)}) = 0$; ii) $d(x)$ is a^{-1}-strongly convex and differentiable.

Associated with d, we define the convex conjugate

$$d^*(z) = \max_{x\in\mathcal{X}}\{\langle z, x\rangle - d(x)\}.$$

According to Danskin's Theorem (Bertsekas, 1999, Proposition 6.1.1), it holds that

$$\nabla d^*(z) = \operatorname*{argmax}_{x\in\mathcal{X}}\{\langle z, x\rangle - d(x)\}.$$

Starting from $x^{(0)}$, CDA produces a sequence of variables $\{x^{(t)}\}_{t\geq 0}$ according to

$$x^{(t)} = \nabla d^*(-z^{(t)}) \tag{3}$$

where

$$z^{(t)} = \sum_{\tau=0}^{t-1} \nabla f(x^{(\tau)}). \tag{4}$$

To investigate the convergence of CDA for nonconvex optimization, we define the following mapping that can be taken as a generalization of the notion of the gradient. The convergence of the proximal gradient descent algorithm (Beck, 2017, Definition 10.5) relies on a similar concept, in the sense that they both represent the change of two consecutive updates. Nevertheless, CDA and the proximal gradient descent generally lead to different trajectories for constrained problems. Therefore, the properties of gradient mapping in CDA need to be re-examined.

Definition 3. (gradient mapping). Suppose that Assumption 1 holds. For any primal-dual pair $(x^{(t)}, z^{(t)})$ generated by (3) and (4), the gradient mapping is defined by

$$G_a(x^{(t)}, z^{(t)}) = \frac{1}{a}\left(\nabla d^*(-z^{(t)}) - \nabla d^*(-z^{(t)} - \nabla f(x^{(t)}))\right). \tag{5}$$

When $\mathcal{X} = \mathbb{R}^m$ and $d(x) = \|x - x^{(0)}\|^2/(2a)$, $G_a(x^{(t)}, z^{(t)}) = \nabla f(x^{(t)})$ for all $t \geq 0$. In this case, clearly, x^* is a stationary point of Problem (1) if and only if there exists z^* such that $G_a(x^*, z^*) = 0$. In the unconstrained case, the relation between $x^{(t)}$ and $z^{(t)}$ is bijective. However, in the presence of constraint, it only holds that (Rockafellar, 1970, Theorem 26.5):

$$-z^{(t)} \in \{\nabla d(x^{(t)})\} \oplus \mathcal{N}_{\mathcal{X}}(x^{(t)}),\ x^{(t)} = \nabla d^*(-z^{(t)}) \quad \forall t \geq 0,$$

where $\oplus$ denotes the Minkowski sum defined by

$$\mathcal{A} \oplus \mathcal{B} := \{a + b | a \in \mathcal{A}, b \in \mathcal{B}\}.$$

Proposition 1. x^* is a stationary point of Problem (1) if and only if there exists some primal-dual pair $(x^*, z^{(t^*)})$ at some $t^* \geq 0$, i.e., $x^* = \nabla d^*(-z^{(t^*)})$, such that $G_a(x^*, z^{(t)}) = 0, \forall t \geq t^*$.

Proof. *Necessity.* Suppose x^* is a stationary point. Pick any z' satisfying $x^* = \nabla d^*(-z')$, and label the time instant as t^*, i.e., $z' = z^{(t^*)}$. By optimality, it holds that

$$-\sum_{\tau=0}^{t^*-1} \nabla f(x^{(\tau)}) - \nabla d(x^*) \in \mathcal{N}_{\mathcal{X}}(x^*).$$

If x^* is a stationary point, we have $-\nabla f(x^*) \in \mathcal{N}_{\mathcal{X}}(x^*)$ and therefore

$$-\sum_{\tau=0}^{t^*} \nabla f(x^{(\tau)}) - \nabla d(x^*) \in \mathcal{N}_{\mathcal{X}}(x^*).$$

This together with the strong convexity of d gives us $x^{(t^*+1)} = x^*$ and $G_a(x^*, z^{(t^*)}) = 0$. By induction, the equality holds for all $t \geq t^*$.

Sufficiency. Suppose there exists some t^* such that $G_a(x^{(t)}, z^{(t)}) = 0, \forall t \geq t^*$. Thus $x^{(t)} = x^{(t+\tau)}, \forall \tau \geq 1$. Denoting $x^{(t)} = x^*, \forall t \geq t^*$, it holds that

$$v - (t - t^*)\nabla f(x^*) \in \mathcal{N}_{\mathcal{X}}(x^*) \text{ and } v \in \mathcal{N}_{\mathcal{X}}(x^*), \forall t \geq t^* + 1$$

where $v = -\sum_{\tau=0}^{t^*-1} \nabla f(x^{(\tau)}) - \nabla d(x^*)$. For the sake of contradiction, suppose

$$-\nabla f(x^*) \notin \mathcal{N}_{\mathcal{X}}(x^*).$$

Then, there must exist some sufficiently large t such that

$$v - (t - t^*)\nabla f(x^*) \notin \mathcal{N}_{\mathcal{X}}(x^*),$$

which yields a contradiction.

Next, we present the convergence rate of CDA for general nonconvex optimization problems.

Theorem 1. Suppose Assumption 1 is satisfied and let $\{x^{(t)}\}_{t\geq 0}$ be the sequence generated by the dual averaging algorithm in (3) and (4) with $a < 2L^{-1}$. Then

i) the sequence $\{f(x^{(t)})\}_{t\geq 0}$ is non-increasing, and $f(x^{(t)}) > \lim_{\tau\to\infty} f(x^{(\tau)})$ if and only if $x^{(t)}$ is not a stationary point;
ii) $G_a(x^{(t)}, z^{(t)}) \to 0$ as $t \to \infty$;
iii) for all $k \geq 1$,

$$\min_{t\leq k} \|G_a(x^{(t)}, z^{(t)})\|^2 \leq \frac{2\left(f(x^{(0)}) - f^*\right)}{a(2 - aL)k}. \tag{6}$$

Remark 1. Theorem 1 provides a sufficient condition for the parameter a, under which CDA converges. In particular, the objective value monotonically decreases until a stationary point is reached, as stated in point i). Furthermore, point ii) indicates that the norm of the suboptimality measure in Definition 3 converges to 0. Finally, point iii) demonstrates that the minimum of squared norm of the measure before arbitrary time $k \geq 1$ is bounded from above by $\mathcal{O}(1/k)$.

Proof of Theorem 1: Before proving Theorem 1, we present Lemma 1, whose proof can be found in (Liu et al., 2022b).

Lemma 1. Suppose Assumption 1 holds. For the sequence $\{x^{(t)}\}_{t\geq 0}$ generated by the dual averaging method in (3) and (4), it holds that

$$\langle \nabla f(x^{(t)}), x^{(t+1)} - x^{(t)} \rangle \leq -\frac{1}{a}\|x^{(t+1)} - x^{(t)}\|^2. \tag{7}$$

We are now in a position to prove Theorem 1.

i) By Assumption 1, we have

$$\begin{aligned} &f(x^{(t+1)}) - f(x^{(t)}) \\ &\leq \langle \nabla f(x^{(t)}), x^{(t+1)} - x^{(t)} \rangle + \frac{L}{2}\|x^{(t+1)} - x^{(t)}\|^2. \end{aligned} \tag{8}$$

Using Lemma 1, we obtain

$$\begin{aligned} f(x^{(t)}) - f(x^{(t+1)}) &\geq \left(\frac{1}{a} - \frac{L}{2}\right)\|x^{(t+1)} - x^{(t)}\|^2 \\ &= \frac{a(2 - aL)}{2}\|G_a(x^{(t)}, z^{(t)})\|^2, \end{aligned} \tag{9}$$

which implies $f(x^{(t)}) \geq f(x^{(t+1)})$. Because the sequence $\{f(x^{(t)})\}_{t\geq 0}$ is non-increasing and bounded from below, it converges. If $x^{(t)}$ is not a stationary point, then $\sum_{\tau=t}^{\infty}\|G_a(x^{(\tau)}, z^{(\tau)})\|^2 \neq 0$ according to Proposition 1, and therefore $f(x^{(t)}) > \lim_{\tau\to\infty} f(x^{(\tau)})$. If $x^{(t)}$ is a stationary point, then $\sum_{\tau=t}^{\infty}\|G_a(x^{(\tau)}, z^{(\tau)})\|^2 = 0$ and $x^{(\tau)} = x^{(t)}, \forall \tau \geq t$, and thus $f(x^{(t)}) = \lim_{\tau\to\infty} f(x^{(\tau)})$.

ii) Because the sequence $\{f(x^{(t)})\}_{t\geq 0}$ converges, $f(x^{(t)}) - f(x^{(t+1)})$ converges to 0 as $t \to \infty$, which in conjunction with (9) gives the desired result.

iii) Summing (9) over $t = 0, 1, \ldots, k$ yields

$$\begin{aligned} &\frac{a(2 - aL)}{2}\sum_{t=0}^{k}\|G_a(x^{(t)}, z^{(t)})\|^2 \leq f(x^{(0)}) - f(x^{(k+1)}) \\ &\leq f(x^{(0)}) - f^* \end{aligned}$$

where the last inequality follows from $f(x^{(k+1)}) \geq f^*$.

4. DISTRIBUTED DUAL AVERAGING FOR NONCONVEX OPTIMIZATION

In this section, we revisit the DDA algorithm in (Liu et al., 2022a), and provide its rate of convergence for nonconvex optimization problems in the form of (1).

The design of DDA is motivated in (Liu et al., 2022a), where the idea is to use dynamic averaging consensus to estimate $z^{(t)}$ in (4) in a distributed manner, followed by a similar step to (3) locally performed by each agent with an inexact version of $z^{(t)}$. The DDA algorithm is detailed in

Algorithm 1 DDA

Input: $a > 0$, a continuously differentiable and a^{-1}-strongly convex proximal function d, $x^{(0)}$

Output: $x_i^{(t)}, t = 1, 2, \ldots$

1: **Initialize:** set $x_i^{(0)} = x^{(0)}$, $z_i^{(0)} = 0$, and $s_i^{(0)} = \nabla f_i(x^{(0)})$ for all $i = 1, \ldots, n$

2: **for** $t = 1, 2, \ldots$, each agent i synchronously **do**

3: collect $z_j^{(t-1)}$ and $s_j^{(t-1)}$ from all agents $j \in \mathcal{N}_i^{(t-1)}$

4: update $z_i^{(t)}$ by

$$z_i^{(t)} = \sum_{j \in \mathcal{N}_i^{(t-1)} \cup \{i\}} p_{ij}^{(t-1)} \left(z_j^{(t-1)} + s_j^{(t-1)} \right)$$

5: compute $x_i^{(t)}$ by

$$x_i^{(t)} = \nabla d_t^*(-z_i^{(t)})$$

6: update $s_i^{(t)}$ by

$$s_i^{(t)} = \sum_{j \in \mathcal{N}_i^{(t-1)} \cup \{i\}} p_{ij}^{(t-1)} s_j^{(t-1)} + \nabla f_i(x_i^{(t)}) - \nabla f_i(x_i^{(t-1)})$$

7: **end for**

Algorithm 1. First, each agent initializes the algorithm by setting the local variables $x_i^{(0)}$, $z_i^{(0)}$, and $s_i^{(0)}$ properly. At each time $t \geq 1$, each agent exchanges the variables $z_i^{(t-1)}$, $s_i^{(t-1)}$ with its neighbors at time $t-1$, and then computes $z_i^{(t)}$, $x_i^{(t)}$, and $s_i^{(t)}$ according to steps 3–6.

Note that Algorithm 1 (Liu et al., 2022a) is different from (Duchi et al., 2011) where nonsmooth optimization problems are considered, and the former introduces an additional variable $s_i^{(t)}$ that is an estimate of $n^{-1} \sum_{i=1}^n \nabla f_i(x_i)$.

4.1 Analysis setup

Similar to (Duchi et al., 2011; Liu et al., 2022a), we construct a sequence of auxiliary variables $\{y^{(t)}\}_{t \geq 1}$ by

$$y^{(t)} = \nabla d^*(-\overline{z}^{(t)}), \tag{10}$$

where $\overline{z}^{(t)} = n^{-1} \sum_{i=1}^n z_i^{(t)}$, and $y^{(0)} = x^{(0)}$. For each $x_i^{(t)}, i = 1, \ldots, n$ and $y^{(t)}$, we have the following relation (Duchi et al., 2011, Lemma 5).

Lemma 2. For every $t \geq 0$ and $i = 1, \ldots, n$, there holds

$$\|x_i^{(t)} - y^{(t)}\| \leq a \|z_i^{(t)} - \overline{z}^{(t)}\|.$$

To proceed, we recall the analysis from (Liu et al., 2022a) in quantifying $\|z_i^{(t)} - \overline{z}^{(t)}\|$. First, we introduce the notations:

$$\mathbf{M} = \begin{bmatrix} \beta & \beta \\ aL(\beta+1) & \beta(aL+1) \end{bmatrix}$$

and

$$\mathbf{x}^{(t)} = \begin{bmatrix} x_1^{(t)} \\ \vdots \\ x_n^{(t)} \end{bmatrix}, \quad \mathbf{y}^{(t)} = \begin{bmatrix} y^{(t)} \\ \vdots \\ y^{(t)} \end{bmatrix}, \quad \overline{g}^{(t)} = \frac{1}{n} \sum_{i=1}^n \nabla f_i(x_i^{(t)}). \tag{11}$$

For the dual variable $\overline{z}^{(t)}$ in (10), one can verify from steps 4 and 6 in Algorithm 1 that

$$\overline{z}^{(t)} = \overline{z}^{(t-1)} + \overline{s}^{(t-1)} = \overline{z}^{(t-1)} + \overline{g}^{(t-1)},$$

where $\overline{s}^{(t)} = n^{-1} \sum_{i=1}^n s_i^{(t)}$. Next, we remark that the update of $\{y^{(t)}\}_{t \geq 1}$ in (10) can be viewed as dual averaging with inexact gradients, whose convergence property is summarized in the following lemma. Its proof can be found in (Liu et al., 2022b), and is omitted here for brevity.

Lemma 3. Suppose Assumption 1 holds. For $y^{(t)}, t = 1, \ldots,$ generated by (10), it holds that $\forall \epsilon > 0$

$$n \left(f(y^{(t)}) - f(y^{(t-1)}) \right) \leq \left(\frac{L+\epsilon}{2} - \frac{1}{a} \right) \|\mathbf{y}^{(t)} - \mathbf{y}^{(t-1)}\|^2 + \frac{L^2}{2\epsilon} \|\mathbf{y}^{(t-1)} - \mathbf{x}^{(t-1)}\|^2. \tag{12}$$

We emphasize that Lemma 3 discerns the behavior of inexact dual averaging in two consecutive updates. This is different from the convex setup where one characterizes the convergence behavior using all the updates in history (Liu et al., 2022a, Lemma 6). We make the change here because, without convexity, the running average of all the updates in history may not be used and the convergence analysis of individual variable is performed.

4.2 Rate of convergence

We begin by defining the consensual stationary point in the distributed case.

Definition 4. (consensual stationary point). A vector $\mathbf{x}^* = [x_1^*; \ldots; x_n^*]$ is called a stationary solution if

$$x_1^* = \cdots = x_n^* \text{ and } -\nabla f(x_i^*) \in \mathcal{N}_{\mathcal{X}}(x_i^*), \forall i = 1, \cdots, n. \tag{13}$$

A sufficient condition to (13) is that there exists a primal-dual pair $(y^*, \overline{z}^{(t^*)})$ at time t^*, i.e., $y^* = \nabla d^*(-\overline{z}^{(t^*)})$, such that

$$n \|G_a(y^*, \overline{z}^{(t)})\|^2 + \sum_{i=1}^n \|x_i^* - y^*\|^2 = 0, \ \forall t \geq t^* \tag{14}$$

where $G_a(\cdot, \cdot)$ is defined in (5). To see this, we note that (14) implies

$$\|G_a(y^*, \overline{z}^{(t)})\|^2 = 0 \quad \text{and} \quad \|x_i^* - y^*\|^2 = 0 \ \forall i = 1, \ldots, n$$

for all $t \geq t^*$, where the former ensures that y^* is a stationary point in the centralized case, and the latter gives $x_i^* = y^*, \forall i = 1, \ldots, n$.

Theorem 2. Suppose Assumptions 1 and 2 are satisfied. If the constant a satisfies

$$\frac{1}{a} > L \cdot \max \left\{ \frac{1}{2} + \frac{4}{3(1-\rho(\mathbf{M}))}, \frac{2\beta}{(1-\beta)^2} \right\},$$

then, it holds that

$$\lim_{t \to +\infty} n \|G_a(y^{(t-1)}, \overline{z}^{(t-1)})\|_{\mathbb{E}}^2 + \|\mathbf{x}^{(t-1)} - \mathbf{y}^{(t-1)}\|_{\mathbb{E}}^2 = 0 \tag{15}$$

and, for all $t \geq 1$,

$$\min_{\tau \leq t} n \|G_a(y^{(\tau-1)}, \overline{z}^{(\tau-1)})\|_{\mathbb{E}}^2 + \|\mathbf{x}^{(\tau-1)} - \mathbf{y}^{(\tau-1)}\|_{\mathbb{E}}^2 \leq \frac{C}{t} \tag{16}$$

where

$$C := \left(\min \left\{ \frac{3L(1-\rho(\mathbf{M}))}{8}, a - \frac{a^2 L}{2} - \frac{4a^2 L}{3(1-\rho(\mathbf{M}))} \right\} \right)^{-1} \times \left(\frac{2\pi^2}{3L(1-\rho(\mathbf{M}))} + n \left(f(y^{(0)}) - f^* \right) \right).$$

Proof of Theorem 2: Before proving Theorem 2, we provide the following lemma. Its proof is similar to the proof of (Liu et al., 2022a, Lemma 5). Due to space limitations, the proof is omitted.

Lemma 4. If Assumption 2 holds and

$$a < \frac{(1-\beta)^2}{2\beta L}, \tag{17}$$

then, for $\mathbf{x}^{(t)}$ and $\mathbf{y}^{(t)}$ defined in (11), it holds that

$$\sum_{\tau=1}^{t}\|\mathbf{x}^{(\tau)}-\mathbf{y}^{(\tau)}\|_{\mathbb{E}}^2 \leq \frac{8}{9(1-\rho(\mathbf{M}))^2}\sum_{\tau=0}^{t-1}\|\mathbf{y}^{(\tau+1)}-\mathbf{y}^{(\tau)}\|_{\mathbb{E}}^2 + \frac{8\pi^2}{9L^2(1-(\rho(\mathbf{M}))^2)}$$

where $\pi^2 = \sum_{i=1}^n \left\|\nabla f_i(x^{(0)}) - \overline{g}^{(0)}\right\|^2$.

Now we are ready to prove Theorem 2.

Recall (12)

$$n\left(f(y^{(t)})-f(y^{(t-1)})\right) \leq \left(\frac{L+\epsilon}{2}-\frac{1}{a}\right)\|\mathbf{y}^{(t)}-\mathbf{y}^{(t-1)}\|^2 + \frac{L^2}{2\epsilon}\|\mathbf{y}^{(t-1)}-\mathbf{x}^{(t-1)}\|^2.$$

Summing it from 1 to t yields

$$n\left(f(y^{(t)})-f(y^{(0)})\right) \leq \left(\frac{L+\epsilon}{2}-\frac{1}{a}\right)\sum_{\tau=1}^{t}\|\mathbf{y}^{(\tau)}-\mathbf{y}^{(\tau-1)}\|^2 + \frac{L^2}{2\epsilon}\sum_{\tau=1}^{t}\|\mathbf{y}^{(\tau-1)}-\mathbf{x}^{(\tau-1)}\|^2.$$

Taking expectation on both sides, we obtain

$$\begin{aligned}
&n\mathbb{E}\left[f(y^{(t)})-f(y^{(0)})\right] \\
&\leq \left(\frac{L+\epsilon}{2}-\frac{1}{a}\right)\sum_{\tau=1}^{t}\|\mathbf{y}^{(\tau)}-\mathbf{y}^{(\tau-1)}\|_{\mathbb{E}}^2 \\
&\quad - \frac{L^2}{2\epsilon}\sum_{\tau=1}^{t}\|\mathbf{y}^{(\tau-1)}-\mathbf{x}^{(\tau-1)}\|_{\mathbb{E}}^2 + \frac{8\pi^2}{9\epsilon(1-(\rho(\mathbf{M}))^2)} \\
&\quad + \frac{8L^2}{9\epsilon(1-\rho(\mathbf{M}))^2}\sum_{\tau=0}^{t-1}\|\mathbf{y}^{(\tau+1)}-\mathbf{y}^{(\tau)}\|_{\mathbb{E}}^2 \\
&= \left(\frac{L+\epsilon}{2}+\frac{8L^2}{9\epsilon(1-\rho(\mathbf{M}))^2}-\frac{1}{a}\right)\sum_{\tau=1}^{t}\|\mathbf{y}^{(\tau)}-\mathbf{y}^{(\tau-1)}\|_{\mathbb{E}}^2 \\
&\quad - \frac{L^2}{2\epsilon}\sum_{\tau=1}^{t}\|\mathbf{y}^{(\tau-1)}-\mathbf{x}^{(\tau-1)}\|_{\mathbb{E}}^2 + \frac{8\pi^2}{9\epsilon(1-\rho(\mathbf{M}))^2}.
\end{aligned}$$

This is equivalent to

$$\begin{aligned}
&na^2\left(\frac{1}{a}-\frac{L+\epsilon}{2}-\frac{8L^2}{9\epsilon(1-\rho(\mathbf{M}))^2}\right)\times \\
&\sum_{\tau=1}^{t}\|G_a(\mathbf{y}^{(\tau-1)},\mathbf{z}^{(\tau-1)})\|^2 + \frac{L^2}{2\epsilon}\sum_{\tau=1}^{t}\|\mathbf{x}^{(\tau-1)}-\mathbf{y}^{(\tau-1)}\|_{\mathbb{E}}^2 \\
&\leq \frac{8\pi^2}{9\epsilon(1-\rho(\mathbf{M}))^2} + n\left(f(y^{(0)})-f^*\right) < +\infty
\end{aligned}$$

because of

$$\|\mathbf{y}^{(\tau)}-\mathbf{y}^{(\tau-1)}\|^2 = na^2\|G_a(\mathbf{y}^{(\tau-1)},\mathbf{z}^{(\tau-1)})\|^2.$$

Thus, (15) holds. Finally, we set $\epsilon = 4L/(3(1-\rho(\mathbf{M}))) > 0$ to obtain (16).

5. NUMERICAL EXAMPLE

Consider the distributed principal component analysis (PCA) problem

$$\min_{\|x\|\leq 1} \; f(x) := -\sum_{i=1}^{n}\|M_i x\|^2$$

where $n = 50$. Each agent i possesses a data matrix $M_i \in \mathbb{R}^{30\times 500}$, where each row $M_i^j, j = 1,\ldots,30$ is randomly generated with zero mean and $\|M_i^j\| \leq 1$. For the communication network among agents, we consider the Bernoulli stochastic network (Kar and Moura, 2008), where a complete graph is taken as the supergraph and at each time t every edge of the set of edges of the supergraph is activated with probability 0.1. Based on it, a Laplacian-based weight matrix (Xiao et al., 2005) is used at each time t.

We initialize Algorithm 1 by randomly generating a 500-dimensional vector with i.i.d. elements drawn from the standard Normal distribution and then projecting it onto the constraint to get $x^{(0)}$. Set the parameter $a = 1$, and $d(x) = \|x - x^{(0)}\|^2/2$ accordingly. We contrast the proposed algorithm with the distributed proximal gradient algorithm (DPGA) in (Jiang et al., 2022) which is able to handle time-varying networks. The stepsize for DPGA is set as $1e{-}4$ in order to stabilize the updates. We remark that DPGA does not have convergence guarantees in stochastic communication networks.

The experiment was repeated 10 times with random seeds. We evaluate the performance of the algorithm via the values of the cost function and the sum of difference in two consecutive updates and consensus error, i.e., $\|\mathbf{x}^{(t)} - \mathbf{x}^{(t-1)}\| + \|\mathbf{x}^{(t)} - \mathbf{1}\otimes\overline{x}^{(t)}\|$, $t \geq 1$ where $\overline{x}^{(t)} = n^{-1}\sum_{i=1}^n x_i^{(t)}$ and $\otimes$ denotes the Kronecker product. We remark that the latter is an approximation of the residual term in Theorem 2. Their mean and standard deviation in 10 runs by the two algorithms are plotted in Figures 1 and 2. Note that a lower value of the cost suggests a closer distance from $\mathbf{x}^{(t)}$ to the principal eigenvector. From the figures, we observe that for DDA both the cost and the residue converge. In addition, the convergence of DDA is faster than PDGA, since the latter has to use a much smaller stepsize to avoid divergence in this experiment. In contrast, DDA remains convergent under a larger range of parameters. This highlights the advantage of DDA in dealing with stochastic communication networks.

6. CONCLUSION

This work examined the convergence rate of dual averaging for nonconvex constrained smooth optimization problems in both centralized and distributed settings. We developed a new suboptimality measure and established its relation to stationarity. The squared norm of such measure converges at rate $\mathcal{O}(1/t)$ for CDA. Under mild conditions on the stochastic communication network, the rate of DDA is proved to be $\mathcal{O}(1/t)$. For future research, we are interested in speeding up DDA for a special class of nonconvex problems satisfying the Kurdyka-Łojasiewicz condition.

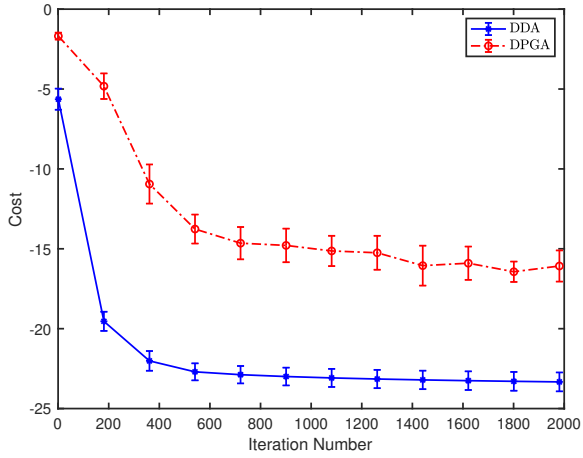

Fig. 1. Convergence of the cost $f(\overline{x}^{(t)})$.

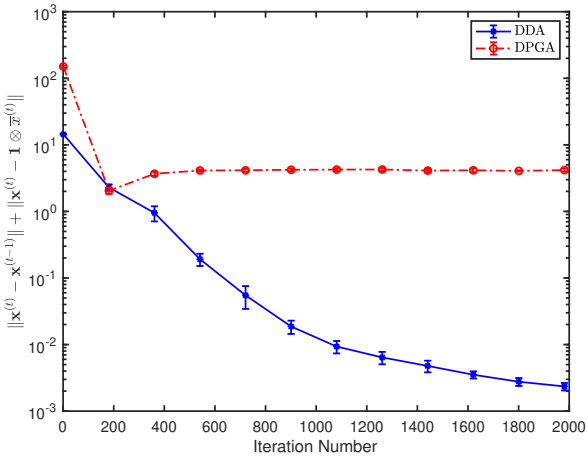

Fig. 2. Convergence of $\|\mathbf{x}^{(t)}-\mathbf{x}^{(t-1)}\|+\|\mathbf{x}^{(t)}-\mathbf{1}\otimes\overline{x}^{(t)}\|$.

REFERENCES

Beck, A. (2017). *First-Order Methods in Optimization*. SIAM.

Bertsekas, D.P. (1999). *Nonlinear Programming*. Athena Scientific.

Boyd, S., Ghosh, A., Prabhakar, B., and Shah, D. (2006). Randomized gossip algorithms. *IEEE Transactions on Information Theory*, 52(6), 2508–2530.

Defazio, A. and Jelassi, S. (2022). Adaptivity without compromise: a momentumized, adaptive, dual averaged gradient method for stochastic optimization. *Journal of Machine Learning Research*, 23, 1–34.

Di Lorenzo, P. and Scutari, G. (2015). Distributed nonconvex optimization over networks. In *2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 229–232. IEEE.

Duchi, J.C., Agarwal, A., and Wainwright, M.J. (2011). Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic control*, 57(3), 592–606.

Fang, H., Harvey, N.J., Portella, V.S., and Friedlander, M.P. (2022). Online mirror descent and dual averaging: Keeping pace in the dynamic case. *Journal of Machine Learning Research*, 23, 1–38.

Héliou, A., Martin, M., Mertikopoulos, P., and Rahier, T. (2020). Online non-convex optimization with imperfect feedback. *Advances in Neural Information Processing Systems*, 33, 17224–17235.

Hong, M., Hajinezhad, D., and Zhao, M.M. (2017). Prox-pda: The proximal primal-dual algorithm for fast distributed nonconvex optimization and learning over networks. In *International Conference on Machine Learning*, 1529–1538. PMLR.

Jiang, X., Zeng, X., Sun, J., and Chen, J. (2022). Distributed proximal gradient algorithm for non-convex optimization over time-varying networks. *IEEE Transactions on Control of Network Systems*.

Kar, S. and Moura, J.M. (2008). Sensor networks with random links: Topology design for distributed consensus. *IEEE Transactions on Signal Processing*, 56(7), 3315–3326.

Lian, X., Zhang, C., Zhang, H., Hsieh, C.J., Zhang, W., and Liu, J. (2017). Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in Neural Information Processing Systems*, 30.

Liu, C., Shi, Y., Li, H., and Du, W. (2022a). Accelerated dual averaging methods for decentralized constrained optimization. *IEEE Transactions on Automatic Control*.

Liu, C., Wu, X., Yi, X., Shi, Y., and Johansson, K.H. (2022b). Rate analysis of dual averaging for nonconvex distributed optimization. *arXiv preprint arXiv:2211.06914*.

Nedic, A., Olshevsky, A., and Shi, W. (2017). Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4), 2597–2633.

Rockafellar, R.T. (1970). *Convex Analysis*. Princeton university press.

Scutari, G. and Sun, Y. (2019). Distributed nonconvex constrained optimization over time-varying digraphs. *Mathematical Programming*, 176(1), 497–544.

Shen, J., Kammara, E.K.H., and Du, L. (2022). Nonconvex, fully distributed optimization based cav platooning control under nonlinear vehicle dynamics. *IEEE Transactions on Intelligent Transportation Systems*.

Xiao, L., Boyd, S., and Lall, S. (2005). A scheme for robust distributed sensor fusion based on average consensus. In *IPSN 2005. Fourth International Symposium on Information Processing in Sensor Networks, 2005.*, 63–70. IEEE.

Xin, R., Das, S., Khan, U.A., and Kar, S. (2021). A stochastic proximal gradient framework for decentralized non-convex composite optimization: Topology-independent sample complexity and communication efficiency. *arXiv preprint arXiv:2110.01594*.

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

Yi, X., Zhang, S., Yang, T., Chai, T., and Johansson, K.H. (2021). Linear convergence of first-and zeroth-order primal–dual algorithms for distributed nonconvex optimization. *IEEE Transactions on Automatic Control*, 67(8), 4194–4201.