Towards an $O(\frac{1}{t})$ convergence rate for distributed dual averaging

Changxin Liu* Huiping Li** Yang Shi*

*Department of Mechanical Engineering, University of Victoria, Victoria, B.C., Canada (e-mail: chxliu@uvic.ca, yshi@uvic.ca). ** School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an, P. R. China (e-mail: lihuiping@nwpu.edu.cn)*

Abstract: Recently, distributed dual averaging has received increasing attention due to its superiority in handling constraints and dynamic networks in multiagent optimization. However, all distributed dual averaging methods reported so far considered nonsmooth problems and have a convergence rate of $O(\frac{1}{\sqrt{t}})$. To achieve an improved convergence guarantee for smooth problems, this work proposes a second-order consensus scheme that assists each agent to locally track the global dual variable more accurately. This new scheme in conjunction with smoothness of the objective ensures that the accumulation of consensus error over time caused by incomplete global information is bounded from above. Then, a rigorous investigation of dual averaging with inexact gradient oracles is carried out to compensate the consensus error and achieve an $O(\frac{1}{t})$ convergence rate. The proposed method is examined in a large-scale LASSO problem.

Keywords: Distributed optimization, smooth optimization, dual averaging, second-order consensus, inexact method.

1. INTRODUCTION

We consider the problem where a team of agents connected via a network manage to optimize the sum of their local interests while respecting certain common constraints. This problem is referred to as distributed optimization, and has been extensively investigated in recent years mainly due to its broad applications. For example, distributed machine learning, formation control of autonomous vehicles, and sensor fusion can be cast as optimization problems of this type. For a recent overview of distributed optimization, please refer to Nedić et al. (2018).

In such a framework, each agent does not have full knowledge of the objective function, therefore it has to communicate with neighbors to estimate the global information, e.g., global gradient or/and mean value of local variables, during the course of optimizer seeking to achieve distributed optimization. Regarding the estimation process, the algorithms in Nedic et al. (2010); Yuan et al. (2016); Liu et al. (2020b) directly seek consensus over local variables based on a doubly stochastic weight matrix, while the optimizer seeking process is guided by the gradient of the local objective. However, due to the fact that local gradients evaluated at global minimizer are not necessarily zero, the two forces caused by consensus and local gradient flows are conflicting with each other, preventing exact optimization when a constant stepsize is used, that is, there always exists a gap between the accumulation point and global minimum. It is worth mentioning that, by using a decaying stepsize, exact optimization may be obtained with, however, a slow $O(\frac{1}{\sqrt{t}})$ convergence rate where t is the time counter. This issue can be solved by an additional

estimation process for the global gradient by using the dynamic average consensus scheme in Zhu and Martínez (2010). It is shown in Varagnolo et al. (2015); Qu and Li (2017) that for unconstrained smooth optimization the algorithm steered by the approximated global gradient obtains exact minimization with an $O(\frac{1}{t})$ rate.

In the methods mentioned above, local estimates about the minimizer are directly generated in the feasible set (in the case of constrained optimization) that is contained in the primal space of variables. There are also some schemes available in the literature where the minimizer seeking process imitates dual methods, e.g., mirror descent in Shahrampour and Jadbabaie (2017) and dual averaging in Duchi et al. (2011). The concept of dual methods was coined by Nemirovsky and Yudin (1983), where a dual model of the objective is updated and a prox-function establishes a mapping from the dual space to the primal to shrink the error bound in primal methods. For example, Duchi et al. (2011) designed a distributed dual averaging algorithm where the global dual variable is gradually learned by a consensus scheme, and demonstrated that minimizing the approximate dual model of the global objective helps bypass the difficulty caused by projection in distributed primal methods. Recent work in Liu et al. (2020a) introduced another averaging step to standard distributed dual averaging to reap a non-ergodic convergence property, which helps deal with distributed optimization problems with coupled constraints. For problems defined over time-varying and unbalanced networks, a distributed dual averaging method with the push-sum technique was reported in Liang et al. (2019).

Although distributed dual methods in the literature have demonstrated advantages over their primal counterparts in terms of constraint handling, convergence rate, and analysis complexity, all the results reported so far focused only on nonsmooth optimization and have a convergence rate of $O(\frac{1}{\sqrt{t}})$. *Considering this, a question naturally arises: If the objective functions exhibit some desired properties, e.g., smoothness, is it possible to accelerate the convergence rate of distributed dual averaging to $O(\frac{1}{t})$?* This work provides affirmative answer to this question. This is made admissible by a new second-order consensus scheme that assists each agent to locally track the global dual variable more accurately. With the new dual estimate, the accumulation of error over time between local primal variables and their mean is proved to admit an upper bound. This together with a rigorous investigation of averaged primal variables yields an accelerated convergence rate.

Notation: $\mathbb{R}$ represents the set of real numbers and $\mathbb{R}^m$ the m-dimensional Euclidean space. In this space, we let $\|\cdot\|_p$ denote the l_p-norm operator, and without specifying p, it stands for the Euclidean norm. Denote the set of non-negative integers by $\mathbb{N}$. Given $k_1, k_2 \in \mathbb{N}$, the notation $\mathbb{N}_{[k_1,k_2]}$ represents the set $\{k \in \mathbb{N} | k_1 \leq k \leq k_2\}$. We denote by $0_m \in \mathbb{R}^m$ and $\mathbf{1}$ the m-dimensional vector of all zeros and all ones, respectively. Given a matrix $P \in \mathbb{R}^{m \times m}$, its spectral radius and singular values are denoted by $\rho(P)$ and $\sigma_1(P) \geq \sigma_2(P) \geq \cdots \geq \sigma_m(P) \geq 0$, respectively.

2. PROBLEM STATEMENT AND PRELIMINARIES

2.1 Problem Statement

Formally, the optimization problem is given by

$$\min_{x \in \mathcal{X}} f(x) = \sum_{i=1}^{n} f_i(x) \qquad (1)$$

where $x \in \mathbb{R}^m$ denotes the global decision variable, $f_i : \mathbb{R}^m \to \mathbb{R}, i \in \mathbb{N}_{[1,n]}$ represents the local objective function that is privately known by agent i, and $\mathcal{X} \subseteq \mathbb{R}^m$ stands for the common constraint set. Throughout the paper, we denote one of the minimizers by x^*. For (1), we make the following standard assumption.

Assumption 1. Each $f_i(x), i \in \mathbb{N}_{[1,n]}$ is convex and has Lipschitz continuous gradient with parameter L, i.e.,

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

The common constraint set $\mathcal{X}$ is convex and closed, and contains the origin.

We use an undirected graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ to describe the communication pattern between agents, where $\mathcal{V} = \{1, \cdots, n\}$ denotes the set of n agents and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ represents the set of channels that connect agents, that is, the pair $(i, j) \in \mathcal{E}$ for $i, j \in \mathcal{V}$ indicates that there exists a link between node i and j. The set of i's neighbors is denoted by $\mathcal{N}_i = \{j \in \mathcal{V} | (j, i) \in \mathcal{E}\}$. The graph is assumed to be fixed and connected in the following.

Assumption 2. The communication graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is fixed and connected.

Based on Assumption 2, a proper weight matrix $P = [p_{ij}]$ can be constructed. In particular, a positive weight p_{ij} is assigned to each communication link $(i, j) \in \mathcal{E}$; for other (i, j) pairs, zero weight is considered. Moreover, the weight matrix satisfies the following assumption.

Assumption 3. 1) P has a strictly positive diagonal, i.e., $p_{ii} > 0$; 2) P is doubly stochastic, i.e., $P\mathbf{1} = \mathbf{1}$ and $\mathbf{1}^{\mathrm{T}} P = \mathbf{1}^{\mathrm{T}}$.

Without loss of generality, we will assume $m = 1$ for ease of notation in the remaining sections, i.e., $\mathbf{1} \otimes I_m = \mathbf{1}$, $P \otimes I_m = P$.

2.2 Preliminaries

Definition 1. A function $d : \mathcal{X} \to \mathbb{R}$ is called a prox-function if 1) $d(x) \geq 0, \forall x \in \mathcal{X}$ and $d(0_m) = 0$; 2) $d(x)$ is differentiable and 1-strongly convex on $\mathcal{X}$, i.e.,

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

Definition 2. For $x, y \in \mathcal{X}$, the Bregman divergence induced by a prox-function d is defined as

$$D_d(x, y) = d(x) - d(y) - \langle \nabla d(y), x - y \rangle.$$

3. ALGORITHM DEVELOPMENT

3.1 Centralized Dual Averaging

This subsection introduces the centralized dual averaging (CDA) Nesterov (2009). CDA generates sequences of the estimates about the minimizer ($\{x_t\}_{t \geq 0}$) and the dual variable ($\{\sum_{k=0}^{t} \nabla f(x_k)\}_{t \geq 0}$) according to the following rule:

$$x_{t+1} = \arg\min_{x \in \mathcal{X}} \Big\{ a_t \sum_{k=0}^{t} \langle \nabla f(x_k), x \rangle + d(x) \Big\} \qquad (2)$$

where $\{a_t\}_{t \geq 0}$ is a sequence of positive control parameters that directly impacts the convergence of CDA. It is shown in Nesterov (2009) that an $O(\frac{1}{\sqrt{t}})$ convergence rate is ensured when a_t decreases at $O(\frac{1}{\sqrt{t}})$ for nonsmooth objective functions. When the objective is smooth, an appropriate constant $a_t = a$ can be used to achieve an $O(\frac{1}{t})$ rate Lu et al. (2018).

For the projection operator in (2), a standard result in convex analysis (Lemma 1 in Nesterov (2009)) is recalled in the following lemma.

Lemma 1. For any $u, v \in \mathbb{R}^m$, we have

$$\| \arg\min_{x \in \mathcal{X}} \{a_t \langle u, x \rangle + d(x)\} - \arg\min_{x \in \mathcal{X}} \{a_t \langle v, x \rangle + d(x)\}\|$$
$$\leq a_t \|u - v\|.$$

3.2 Design of A New Distributed Dual Averaging Scheme

In the literature, several distributed dual averaging algorithms have been developed accounting for different communication patterns among agents. Generally speaking, they both involve iteratively estimating the global dual variable $\sum_{k=0}^{t} \nabla f(x_k)$ in (2) in the following way:

$$q_{i,t+1} = \sum_{j=1}^{n} p_{ij} q_{j,t} + \nabla f_i(x_{i,t+1})$$

where $q_{i,t+1}$ is an estimate of $\sum_{k=0}^{t} \nabla f(x_k)$ locally maintained by agent i at time t, and $x_{i,t}$ is local estimate about

the global minimizer. However, it is shown in Liu et al. (2020b) that $q_{i,t}$ does not necessarily converge to the dual variable. Therefore, the control sequence $\{a_t\}_{t\geq 0}$ has to be decreasing for a slow convergence rate, i.e., $O(\frac{1}{\sqrt{t}})$.

To possibly accelerate convergence using a constant control sequence, the global dual variable must be more accurately estimated. Motivated by this, we propose to track the global dual variable according to the following rule:

$$s_{i,t+1} = \sum_{j=1}^{n} p_{ij}s_{j,t} + \triangledown f_i(x_{i,t+1}) - \triangledown f_i(x_{i,t}) \qquad (3a)$$

$$h_{i,t+1} = \sum_{j=1}^{n} p_{ij}h_{j,t} + s_{i,t+1} - s_{i,t}. \qquad (3b)$$

Note that when $h_{i,0} = s_{i,0}$ one gets the new dual estimate

$$\sum_{k=0}^{t+1} h_{i,k} = \sum_{j=1}^{n} p_{ij} \sum_{k=0}^{t} h_{j,k} + s_{i,t+1}. \qquad (4)$$

Thanks to it, the estimate about the global minimizer can be generated as follows.

$$x_{i,t+1} = \arg\min_{x\in\mathcal{X}}\left\{\sum_{k=0}^{t} a\langle h_{i,k}, x\rangle + d(x)\right\}. \qquad (5)$$

Denote $\bar{h}_t = \frac{1}{n}\sum_{i=1}^{n} h_{i,t}$, $\bar{s}_t = \frac{1}{n}\sum_{i=1}^{n} s_{i,t}$, and $g(t) = \frac{1}{n}\sum_{i=1}^{n} \triangledown f_i(x_{i,t})$. The following conservation property holds true.

Lemma 2. If $h_{i,0} = s_{i,0} = \triangledown f_i(x_{i,0}), i \in \mathcal{V}$, then

$$\bar{h}_{t+1} = \bar{s}_{t+1} = g_{t+1}.$$

Proof. The proof follows from projecting (3) into the average space.

The proposed algorithm is summarized in the following.

Initialization: Set $t = 0$, $x_{i,0} = \arg\min_{x\in\mathcal{X}} d(x) = 0_m$, $h_{i,0} = s_{i,0} = \triangledown f_i(x_{i,0}), \forall i \in \mathcal{V}$. Choose a constant control sequence $a_t = a$.

Each agent $i \in \mathcal{V}$ (in parallel)

1) Receives $s_{j,t}, h_{j,t}, \forall j \in \mathcal{N}_i$;

2) Performs local computation in (3) and (5);

3) Broadcasts $s_{i,t+1}, h_{i,t+1}$ to $j \in \mathcal{N}_i$;

4) Sets $t = t + 1$.

4. MAIN RESULT

First, we set up an auxiliary sequence $\{y_t\}_{t\geq 0}$ that evolves according to the following rule

$$y_{t+1} = \arg\min_{x\in\mathcal{X}}\left\{\sum_{k=0}^{t} a\langle g_k, x\rangle + d(x)\right\}, \qquad (6)$$

where the initial vector $y_0 = \arg\min_{x\in\mathcal{X}} d(x) = 0_m$. Then, the deviation between $\{x_{i,t}\}_{t\geq 0}$ and $\{y_t\}_{t\geq 0}$ is analyzed. Finally, the convergence of $\{y_t\}_{t\geq 0}$ to the global minimizer is shown.

Define

$$\mathbf{x}_t = \begin{bmatrix} x_{1,t} \\ x_{2,t} \\ \vdots \\ x_{n,t} \end{bmatrix}, \mathbf{h}_t = \begin{bmatrix} h_{1,t} \\ h_{2,t} \\ \vdots \\ h_{n,t} \end{bmatrix}, \mathbf{s}_t = \begin{bmatrix} s_{1,t} \\ s_{2,t} \\ \vdots \\ s_{n,t} \end{bmatrix}, \nabla_t = \begin{bmatrix} \nabla f_1(x_{1,t}) \\ \nabla f_2(x_{2,t}) \\ \vdots \\ \nabla f_n(x_{n,t}) \end{bmatrix}.$$

and $\mathbf{z}_{k+1} = \sum_{l=0}^{k} a\mathbf{h}_l$.

The following lemma establishes the relation between sequences $\{x_{i,t}\}_{t\geq 0}$ and $\{y_t\}_{t\geq 0}$; the deviation between them represents the consensus error to be compensated in convergence rate analysis.

Lemma 1. For

$$E(a) = \begin{bmatrix} \beta & a \\ L(\beta+1) & \beta + La \end{bmatrix},$$

where $\beta = \sigma_2(P)$, if $\rho(E(a)) < 1$, it holds that

$$\sum_{k=0}^{t-1} \|\mathbf{x}_k - \mathbf{1}y_k\|^2 \leq \frac{n}{\left(1 - \rho(E(a))\right)^2} \sum_{j=0}^{t-1} \|y_{j+1} - y_j\|^2. \qquad (7)$$

Proof. Please refer to Appendix A.

The following lemma plays a similar role with the well-known dual averaging inequality (Theorem 2 in Nesterov (2009)) for nonsmooth optimization in convergence analysis. However, it further makes use of the smoothness of the objective in order to provide a much tighter bound for a faster convergence rate.

Lemma 2. For $\{y_t\}_{t\geq 0}$ generated by (6), it holds

$$\sum_{k=0}^{t-1} \langle ag_k, y_{k+1} - x^*\rangle \leq d(x^*) - \sum_{k=0}^{t-1} D_d(y_{k+1} - y_k). \qquad (8)$$

Proof. The proof is postponed to Appendix B.

We are now in a position to present the main result.
Theorem 3. If $\rho(E(a)) < 1$ and

$$aL + \frac{aL}{\left(1 - \rho(E(a))\right)^2} \leq \frac{1}{2},$$

then

$$f(\tilde{y}_t) - f(x^*) \leq \frac{nd(x^*)}{at}, \qquad (9)$$

where $\tilde{y}_t = \frac{1}{t}\sum_{k=0}^{t-1} y_{k+1}$.

Proof. Consider

$$\sum_{j=1}^{n} a\Big(f_j(y_{k+1}) - f_j(x^*) \Big)$$

$$\leq \sum_{j=1}^{n} a\Big(\frac{L}{2} \|y_{k+1} - x_{j,k}\|^2$$

$$+ f_j(x_{j,k}) + \langle \nabla f_j(x_{j,k}), y_{k+1} - x_{j,k} \rangle - f_j(x^*) \Big)$$

$$\leq \sum_{j=1}^{n} a\Big(\frac{L}{2} \|y_{k+1} - x_{j,k}\|^2 + \langle \nabla f_j(x_{j,k}), y_{k+1} - x^* \rangle \Big)$$

$$= \sum_{j=1}^{n} a\Big(\frac{L}{2} \|y_{k+1} - x_{j,k}\|^2 \Big) + n\langle ag_k, y_{k+1} - x^* \rangle$$

$$\leq \frac{a}{2} L \|\mathbf{1} y_{k+1} - \mathbf{x}_k\|^2 + n\langle ag_k, y_{k+1} - x^* \rangle$$

$$\leq \frac{a}{2} L \|\mathbf{1} y_{k+1} - \mathbf{1} y_k + \mathbf{1} y_k - \mathbf{x}_k\|^2 + n\langle ag_k, y_{k+1} - x^* \rangle$$

$$\leq aL\Big(n\|y_{k+1} - y_k\|^2 + \|\mathbf{1} y_k - \mathbf{x}_k\|^2 \Big) + n\langle ag_k, y_{k+1} - x^* \rangle,$$
(10)

where the first inequality follows from the use of Lipschitz continuity of the gradient.

This together with convexity of f_j allows us to further get

$$at\Big(f(\tilde{y}_t) - f(x^*) \Big) \leq at\Big(\sum_{j=1}^{n} f_j(\tilde{y}_t) - f(x^*) \Big)$$

$$\leq \sum_{k=0}^{t-1} \sum_{j=1}^{n} a\Big(f_j(y_{k+1}) - f_j(x^*) \Big)$$

$$\leq aLn \sum_{k=0}^{t-1} \|y_{k+1} - y_k\|^2 + n \sum_{k=0}^{t-1} \langle ag_k, y_{k+1} - x^* \rangle$$

$$+ aL \sum_{k=0}^{t-1} \|\mathbf{1} y_k - \mathbf{x}_k\|^2$$

$$\leq \Big(aLn + \frac{aLn}{\big(1 - \rho(E(a))\big)^2} \Big) \sum_{k=0}^{t-1} \|y_{k+1} - y_k\|^2$$

$$+ nd(x^*) - n \sum_{k=0}^{t-1} D_d(y_{k+1} - y_k).$$

Due to $D_d(y_{k+1}, y_k) \geq \frac{1}{2} \|y_{k+1} - y_k\|^2$, we arrive at

$$at\Big(f(\tilde{y}_t) - f(x^*) \Big)$$

$$\leq n\Big(aL + \frac{aL}{\big(1 - \rho(E(a))\big)^2} - \frac{1}{2} \Big) \sum_{k=0}^{t-1} \|y_{k+1} - y_k\|^2 + nd(x^*),$$
(11)

thereby completing the proof.

Remark 1. Theorem 3 states that $\tilde{y}_t$ converges to the global minimizer at an $O(\frac{1}{t})$ rate. By (7) and convexity of the 2-norm operator, one has

$$t\|\tilde{\mathbf{x}}_t - \mathbf{1}\tilde{y}_t\|^2 \leq \sum_{k=1}^{t} \|\mathbf{x}_k - \mathbf{1} y_k\|^2$$

$$\leq \frac{n}{\big(1 - \rho(E(a))\big)^2} \sum_{j=0}^{t} \|y_{j+1} - y_j\|^2,$$
(12)

where $\tilde{\mathbf{x}}_t = \frac{1}{t} \sum_{k=0}^{t-1} \mathbf{x}_{k+1}$. Moreover, from (11), we know that the right-hand side of (12) remains finite as t ap-

proaches infinity. Therefore, $\|\tilde{x}_{i,t} - \tilde{y}_t\|^2$ converges at an $O(\frac{1}{t})$ rate, where $\tilde{x}_{i,t} = \frac{1}{t} \sum_{k=0}^{t-1} x_{i,k+1}$. This implies that $\tilde{x}_{i,t}$ shares a similar convergence guarantee with $\tilde{y}_t$.

5. SIMULATION

To verify the proposed method, we apply it to a large-scale LASSO problem. In this problem, the data tuple (y_i, A_i) available at each agent $i \in \mathcal{V}$ satisfies the following equation:

$$y_i = A_i x^* + b_i,$$

where $A_i \in \mathbb{R}^{p_i \times m}$, $y_i \in \mathbb{R}^{p_i}$, and $b_i \in \mathbb{R}^m$ is the additive Gaussian noise with zero mean and variance σ^2. Usually, $p_i \ll m$ and x^* is sparse. To recover x^*, the following distributed optimization problem is considered:

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

In the simulation, we set $n = 50$, $m = 10000$, $p_i = 20, \forall i \in \mathcal{V}$. The matrix A_i is randomly generated with $\mathcal{N}(0,1)$ elements. The minimizer x^* is a sparse vector that only has 50 non-zero $\mathcal{N}(0,1)$ entries. The variance for noise $b_i, \forall i \in \mathcal{V}$ is set as $\sigma^2 = 0.01$. Set $R = 1.1 * \|x^*\|_1$. The communication network is characterized by an Erdos-Renyi graph with a 0.1 connectivity ratio, and the doubly stochastic matrix P associated with the graph is derived by following the Metropolis-Hastings rule.

For the purpose of comparison, the distributed projected gradient method (DPG) in Nedic et al. (2010), and the distributed dual averaging (DDA) in Duchi et al. (2011) are simulated. To accommodate the theoretical results developed therein, the stepsize for DPG is chosen as $\frac{1}{\sqrt{t}}$; the control sequence in DDA is set as $a_t = \frac{1}{\sqrt{t}}$. The control sequence for the proposed new DDA (N-DDA) is set as $a = \frac{1}{m}$. The initial primal variable for DPG is set as $x_{i,0} = 0, \forall i \in \mathcal{V}$.

The simulation results are reported in the following. The performance is evaluated in terms of two criteria, that is, the primal variable residual of the first agent, i.e., $\frac{\|x_{1,t} - x^*\|^2}{\|x^*\|^2}$, and the objective value over the number of local iteration times. The results suggest that the proposed N-DDA enjoys a faster convergence rate. This is compatible with the theoretical results that DPG and DDA have a rate of $O(\frac{1}{\sqrt{t}})$ while N-DDA converges at an $O(\frac{1}{t})$ rate.

6. CONCLUSION

In this work, we proposed a new distributed dual averaging method tailored for smooth problems that has a convergence rate of $O(\frac{1}{t})$. This is made possible by a second-order consensus scheme that provides an accurate local estimate of the dual variable and a new analysis framework for dual averaging with inexact gradients. This work opens several avenues for future research, including the extension to smooth and strongly convex problems, and dynamic communication networks.

REFERENCES

Duchi, J.C., Agarwal, A., and Wainwright, M.J. (2011). Dual averaging for distributed optimization: Conver-
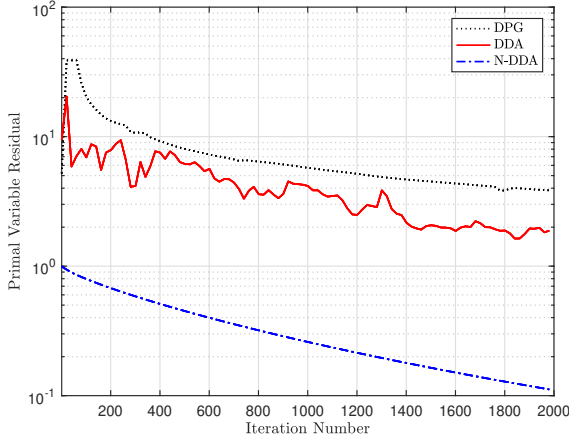
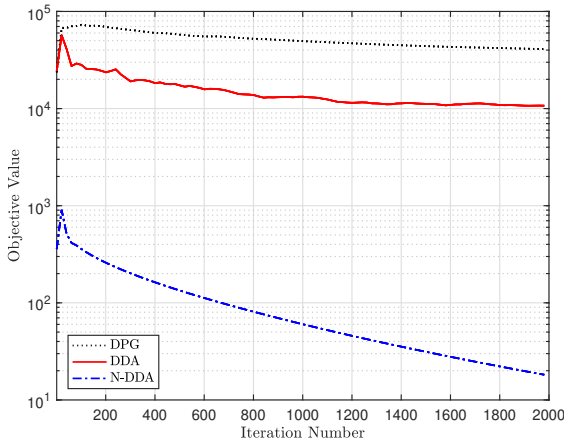Fig. 1. Convergence of the primal variable residual.

Fig. 2. Convergence of the objective value.

gence analysis and network scaling. *IEEE Transactions on Automatic control*, 57(3), 592–606.

Liang, S., Yin, G., et al. (2019). Dual averaging push for distributed convex optimization over time-varying directed graphs. *IEEE Transactions on Automatic Control*.

Liu, C., Li, H., and Shi, Y. (2020a). A unitary distributed subgradient method for multi-agent optimization with different coupling sources. *Automatica*, 114, 108834.

Liu, C., Li, H., Shi, Y., and Xu, D. (2020b). Distributed event-triggered gradient method for constrained convex minimization. *IEEE Transactions on Automatic Control*, 65(2), 778–785.

Lu, H., Freund, R.M., and Nesterov, Y. (2018). Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1), 333–354.

Nedić, A., Olshevsky, A., and Rabbat, M.G. (2018). Network topology and communication-computation trade-offs in decentralized optimization. *Proceedings of the IEEE*, 106(5), 953–976.

Nedic, A., Ozdaglar, A., and Parrilo, P.A. (2010). Constrained consensus and optimization in multi-agent networks. *IEEE Transactions on Automatic Control*, 55(4), 922–938.

Nemirovsky, A.S. and Yudin, D.B. (1983). Problem complexity and method efficiency in optimization.

Nesterov, Y. (2009). Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1), 221–259.

Qu, G. and Li, N. (2017). Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3), 1245–1260.

Shahrampour, S. and Jadbabaie, A. (2017). Distributed online optimization in dynamic environments using mirror descent. *IEEE Transactions on Automatic Control*, 63(3), 714–725.

Varagnolo, D., Zanella, F., Cenedese, A., Pillonetto, G., and Schenato, L. (2015). Newton-raphson consensus for distributed convex optimization. *IEEE Transactions on Automatic Control*, 61(4), 994–1009.

Williams, K.S. (1992). The n th power of a 2×2 matrix. *Mathematics Magazine*, 65(5), 336–336.

Yuan, K., Ling, Q., and Yin, W. (2016). On the convergence of decentralized gradient descent. *SIAM Journal on Optimization*, 26(3), 1835–1854.

Zhu, M. and Martínez, S. (2010). Discrete-time dynamic average consensus. *Automatica*, 46(2), 322–329.

Appendix A. PROOF OF LEMMA 1

Since $\mathbf{s}_0 = \mathbf{h}_0 = \nabla_0$, from (3) we have

$$\mathbf{z}_k = P\mathbf{z}_{k-1} + a\mathbf{s}_{k-1}.$$

By subtracting $\sum_{l=0}^{k-1} ag_l$ on both sides and the triangle inequality, we get

$$\|\mathbf{z}_k - \mathbf{1}\sum_{l=0}^{k-1} ag_l\|$$
$$\leq \|P\mathbf{z}_{k-1} - \mathbf{1}\sum_{l=0}^{k-2} ag_l\| + a\|\mathbf{s}_{k-1} - \mathbf{1}g_{k-1}\| \quad (A.1)$$
$$\leq \beta\|\mathbf{z}_{k-1} - \mathbf{1}\sum_{l=0}^{k-2} ag_l\| + a\|\mathbf{s}_{k-1} - \mathbf{1}g_{k-1}\|.$$

Similarly, it holds that

$$\|\mathbf{s}_k - \mathbf{1}g_k\|$$
$$= \|P\mathbf{s}_{k-1} - \mathbf{1}g_{k-1} + \nabla_k - \nabla_{k-1} - \mathbf{1}g_k + \mathbf{1}g_{k-1}\| \quad (A.2)$$
$$\leq \beta\|\mathbf{s}_{k-1} - \mathbf{1}g_{k-1}\| + L\|\mathbf{x}_k - \mathbf{x}_{k-1}\|,$$

where the fact

$$\|\nabla_k - \nabla_{k-1} - \mathbf{1}g_k + \mathbf{1}g_{k-1}\| \leq \|\nabla_k - \nabla_{k-1}\|$$

and the Lipschitz continuity of the gradient are used to get the last inequality. Using Lemma 1 over $\|\mathbf{x}_k - \mathbf{1}y_k\|$ and $\|\mathbf{x}_{k-1} - \mathbf{1}y_{k-1}\|$, and (A.1) allows us to further get

$$\|\mathbf{s}_k - \mathbf{1}g_k\|$$
$$\leq \beta\|\mathbf{s}_{k-1} - \mathbf{1}g_{k-1}\| + L\|\mathbf{x}_k - \mathbf{1}y_k\| + L\|\mathbf{x}_{k-1} - \mathbf{1}y_{k-1}\|$$
$$\quad + L\|\mathbf{1}y_k - \mathbf{1}y_{k-1}\|$$
$$\leq (\beta + La)\|\mathbf{s}_{k-1} - \mathbf{1}g_{k-1}\| + \sqrt{n}L\|y_k - y_{k-1}\|$$
$$\quad + (L + \beta L)\|\mathbf{z}_{k-1} - \mathbf{1}\sum_{l=0}^{k-2} ag_l\|.$$
$$(A.3)$$

From (A.1) and (A.3), the following linear system inequality can be established:

$$\begin{bmatrix} \|\mathbf{z}_k - \mathbf{1}\sum_{l=0}^{k-1} ag_l\| \\ \|\mathbf{s}_k - \mathbf{1}g_k\| \end{bmatrix} \le E(a)\begin{bmatrix} \|\mathbf{z}_{k-1} - \mathbf{1}\sum_{l=0}^{k-2} ag_l\| \\ \|\mathbf{s}_{k-1} - \mathbf{1}g_{k-1}\| \end{bmatrix}$$
$$+ \sqrt{n}L\begin{bmatrix} 0 \\ \|y_k - y_{k-1}\| \end{bmatrix}. \quad (A.4)$$

Since $\mathbf{h}_0 = \mathbf{s}_0 = \nabla_0 = \mathbf{1}g_0$ by initialization, it holds that

$$\begin{bmatrix} \|\mathbf{z}_k - \mathbf{1}\sum_{l=0}^{k-1} ag_l\| \\ \|\mathbf{s}_k - \mathbf{1}g_k\| \end{bmatrix} \le \sqrt{n}L\sum_{j=0}^{k-1}(E(a))^{k-j-1}\begin{bmatrix} 0 \\ \|y_{j+1} - y_j\| \end{bmatrix}. \quad (A.5)$$

It is easy to check that the eigenvalues of $E(a)$ are
$$\frac{2\beta + aL \pm \sqrt{a^2 L^2 + 4(\beta+1)aL}}{2}.$$
Since $\rho(E(a)) < 1$, one readily has $aL < \beta + 1$. Then, according to Williams (1992),

$$\|\mathbf{z}_k - \mathbf{1}\sum_{l=0}^{k-1} ag_l\|$$
$$\le \frac{\sqrt{n}aL}{\sqrt{a^2 L^2 + 4(\beta+1)aL}}\sum_{j=0}^{k-1}(\lambda_1^{k-j-1} - \lambda_2^{k-j-1})\|y_{j+1} - y_j\|$$
$$\le \sqrt{n}\sum_{j=0}^{k-1}\rho(E(a))^{k-j-1}\|y_{j+1} - y_j\|, \quad (A.6)$$

where $\lambda_1 > \lambda_2$ are eigenvalues of $E(a)$. Therefore

$$\sum_{k=0}^{t-1}\|\mathbf{z}_k - \mathbf{1}\sum_{l=0}^{k-1} ag_l\|^2$$
$$\le n\sum_{k=1}^{t-1}\Big(\sum_{j=0}^{k-1}\rho(E(a))^{k-j-1}\|y_{j+1} - y_j\|\Big)^2$$
$$\le n\sum_{k=1}^{t-1}$$
$$\Big(\sum_{j=0}^{k-1}\big(\rho(E(a))^{\frac{k-j-1}{2}}\big)^2\sum_{j=0}^{k-1}\big(\rho(E(a))^{\frac{k-j-1}{2}}\|y_{j+1}-y_j\|\big)^2$$
$$\le n\sum_{k=1}^{t-1}\frac{1}{1-\rho(E(a))}\sum_{j=0}^{k-1}\rho(E(a))^{k-j-1}\|y_{j+1}-y_j\|^2$$
$$= \frac{n}{1-\rho(E(a))}\sum_{k=1}^{t-1}\sum_{j=0}^{k-1}\rho(E(a))^{k-j-1}\|y_{j+1}-y_j\|^2$$
$$\le \frac{n}{(1-\rho(E(a)))^2}\sum_{j=0}^{t-1}\|y_{j+1}-y_j\|^2,$$

which together with Lemma 1 yields (7).

Appendix B. PROOF OF LEMMA 2

Define
$$m_k(x) = \langle\sum_{l=0}^{k} ag_l, x\rangle + d(x).$$
We then have
$$m_k(x) = m_{k-1}(x) + \langle ag_k, x\rangle$$

According to the definition of Bregman divergence, we have
$$D_{m_{k-1}}(y_{k+1} - y_k)$$
$$= m_{k-1}(y_{k+1}) - m_{k-1}(y_k) - \langle\nabla m_{k-1}(y_k), y_{k+1} - y_k\rangle$$
which is equivalent to
$$D_d(y_{k+1} - y_k)$$
$$= m_{k-1}(y_{k+1}) - m_{k-1}(y_k) - \langle\nabla m_{k-1}(y_k), y_{k+1} - y_k\rangle.$$
Since
$$y_k = \arg\min_{x\in\mathcal{X}} m_{k-1}(x),$$
by the optimality condition we have
$$\langle\nabla m_{k-1}(y_k), y_{k+1} - y_k\rangle \ge 0$$
and therefore
$$0 \le m_{k-1}(y_{k+1}) - m_{k-1}(y_k) - D_d(y_{k+1} - y_k)$$
$$= m_k(y_{k+1}) - \langle ag_k, y_{k+1}\rangle - m_{k-1}(y_k) - D_d(y_{k+1} - y_k)$$
which is equivalent to
$$\langle ag_k, y_{k+1}\rangle \le m_k(y_{k+1}) - m_{k-1}(y_k) - D_d(y_{k+1} - y_k).$$
Summing the above equation over k from 0 to $t-1$ yields
$$\sum_{k=0}^{t-1}\langle ag_k, y_{k+1}\rangle$$
$$\le m_{t-1}(y_t) - m_0(y_1) + \langle ag_0, y_1\rangle - \sum_{k=0}^{t-1}D_d(y_{k+1} - y_k)$$
$$= m_{t-1}(y_t) - \sum_{k=0}^{t-1}D_d(y_{k+1} - y_k) \quad (B.1)$$

We turn to consider
$$\sum_{k=0}^{t-1}\langle ag_k, -x^*\rangle \le \max_{x\in\mathcal{X}}\Big\{\langle\sum_{k=0}^{t-1}ag_k, -x\rangle - d(x)\Big\} + d(x^*)$$
$$= -\min_{x\in\mathcal{X}}\Big\{\langle\sum_{k=0}^{t-1}ag_k, x\rangle + d(x)\Big\} + d(x^*)$$
$$= -m_{t-1}(y_t) + d(x^*),$$
which in conjunction with (B.1) gives rise to the inequality in (8).