Neurocomputing 177 (2016) 643–650

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

# Distributed mirror descent method for multi-agent optimization with delay ☆

Jueyou Li [a,b], Guo Chen [b], Zhaoyang Dong [b], Zhiyou Wu [a,*]

[a] School of Mathematical Science, Chongqing Normal University, Chongqing 400047, China
[b] School of Electrical and Information Engineering, The University of Sydney, NSW 2006, Australia

A R T I C L E  I N F O

*Article history:*
Received 28 September 2015
Received in revised form
23 November 2015
Accepted 2 December 2015
Communicated by Hongyi Li
Available online 17 December 2015

*Keywords:*
Delayed gradient
Multi-agent network
Distributed algorithm
Mirror descent
Convex optimization

A B S T R A C T

This paper investigates a distributed optimization problem associated a time-varying multi-agent network with the presence of delays, where each agent has local access to its convex objective function, and cooperatively minimizes a sum of convex objective functions of the agents over the network. Based on the mirror descent method, we develop a distributed algorithm to solve this problem by exploring the delayed gradient information. Furthermore, we analyze the effects of delayed gradients on the convergence of the algorithm and provide an explicit bound on the convergence rate as a function of the delay parameter, the network size and topology. Our results show that the delays are asymptotically negligible for smooth problems. The proposed algorithm can be viewed as a generalization of the distributed gradient-based projection methods since it utilizes a customized Bregman divergence instead of the usual Euclidean squared distance. Finally, some simulation results on a logistic regression problem are presented to demonstrate the effectiveness of the algorithm.



## 1. Introduction

Distributed optimization over a multi-agent network has attracted considerable recent interest in many applications. This is mainly motivated by the emergence of large-scale networks such as Internet networks, mobile ad hoc networks and wireless sensor networks, characterized by the lack of centralized access to information and time-varying connectivity. Therefore, optimization algorithms embedded in the underling multi-agent network should be completely distributed with local agent communications and adaptive to the changes in the network topology. Starting with the pioneering work [1–3], there has been extensively studies on solving optimization problems in a distributed manner [4–6]. For more details, we can also refer to the recent book [7] and references therein. Recently, by utilizing the idea of consensus averaging over a network [8,9], many researchers have investigated various multi-agent optimization problems arising in the engineering community [10–17]. In [18], Li et al. designed a novel observer-based adaptive sliding mode control approach of nonlinear Markovian jump systems with partly unknown transition probabilities and improved the performance of the considered systems. In [19,20], Zhou et al. developed decentralized adaptive control schemes for robot finger dynamics and MIMO nonlinear systems with input saturation. In [21–23], the authors proposed consensus control methods based on event-triggering strategy over multi-agent network systems.

In the literature several useful algorithms have been developed for the distributed optimization. Of particular interest is the distributed gradient-based projection (DGP) methods, see, e.g., [24–28]. Based on the consensus strategy, Nedić et al. [6] originally proposed a distributed (sub)gradient method that involves every agent minimizing its own objective while exchanging information locally with other agents over a network. Their contribution is that the explicit error bounds on the convergence rate of the algorithm have been established. However, there only unconstrained optimization is considered. Later, the authors in [24] extended the results of [6] to a constrained case, where the estimates of each agent are restricted to lie in different constrained sets. Duchi et al. in [29] studied a constrained distributed optimization based on dual (sub)gradient averaging [8]. The contribution of that paper is to show that the number of iterations required by their proposed algorithm scales inversely in the spectral gap of the network. In [30,31], the authors developed distributed primal-dual (sub)gradient methods for solving a class of distributed problems with

☆This research was partially supported by the NSFC (11501070, 71501017 and 11471062), by the Chinese Scholarship Council Grant 201508500038, by the Natural Science Foundation Projection of Chongqing (cstc2015jcyjA0382, cstc2013jjB00001 and cstc2013jcyjA00029), by the Chongqing Municipal Education Commission under Grant KJ1500301 and by the Chongqing Normal University Research Foundation 15XLB005.

* Corresponding author.
*E-mail addresses:* lijueyou@163.com (J. Li), guo.chen@sydney.edu.au (G. Chen), joe.dong@sydney.edu.au (Z. Dong), zywu@cqnu.edu.cn (Z. Wu).

http://dx.doi.org/10.1016/j.neucom.2015.12.017
0925-2312/

equality or inequality constraints, by characterizing the primal and dual optimal solutions as the saddle points of the Lagrangian of the problem under consideration.

Due to requiring only the calculation of (sub)gradients and projections, the DGP methods have been widely used in large-scale distributed optimization [5]. However, it is sometimes challenging for DGP methods to generate projections for certain objective functions and constraint sets yielding inefficient updates [32]. One of typical examples is the entropy-based loss function with the constraint set being the unit simplex. To overcome the difficulty, Xi et. al [33] recently proposed a novel distributed algorithm depending on the mirror descent method [34], which employs the Bregman distance instead of the Euclidean squared distance. Compared with DGP methods, the primary advantage of their proposed algorithm can efficiently deal with certain non-Euclidean projections and improve the performance in high dimensions.

The distributed algorithms presented in earlier works [24,29,30,33] assume that at any time, each agent has access to estimate the states of its immediate neighbors. As pointed out in [7], this may be impossible in communication networks where delays exist in the transmission of agent estimates over a communication channel. A natural extension is to study an asynchronous operation of optimization algorithms in the presence of delays. In [35–37], the authors generalized the DGP methods or the distributed dual averaging (sub)gradient methods to handle distributed optimization over a network with communication delays or quantized communication. In addition, the convergence rates of distributed algorithms based on delayed gradient information were analyzed in [38,39]. Recently, various types of delays over networks have been extensively investigated by many researchers, for examples, [40–46].

As it is well-known, establishing the rate properties of distributed algorithms in multi-agent systems is essential in understanding the robustness of the system against dynamic changes. However, the convergence analysis of the algorithms is more challenging due to delays. In this paper we will extend the distributed mirror descent (DMD) method to handle with the setting, where communication delays over a time-varying multi-agent network are considered.

The work in this paper is closely related to the previous works [24,33,39]. Our algorithm is based on the initial discovery [33], but we consider the presence of communication delays in the DMD method and investigate the effect of delays on the convergence properties of the algorithm. Therefore, the proposed algorithm in this paper can be viewed as a generalization of the DMD algorithm. In particular, we also extend the DGP methods [24,25] to the case with communication delays, since the Euclidean projection in the DGP methods is replaced by the more general Bregman projection. Also related is the delayed distributed dual averaging (DDDA) algorithm presented in [39] by extending the distributed dual averaging algorithm to the case, where only the delayed gradient information is available over the network. However, our algorithm essentially differs from the algorithm described in [39], where the proximal projection generating from a proximal function is used in the DDDA algorithm rather than the Bregman projection generating from a Bregman divergence function in our algorithm, although both algorithms take into gradient delays consideration for solving distributed optimization. In addition, the difference lies in that the DDDA algorithm makes a global approximation while our algorithm only makes a local approximation.

There are two main contributions in this paper. The first contribution is the extension of the distributed mirror descent algorithm to the delayed case, where the procedure receives out-of-date gradients instead of current gradients. The second main contribution is a careful analysis that demonstrates the effects of delays on the convergence rate of the proposed algorithm. We prove that the algorithm can asymptotically converge to the optimal solution as the iteration approaches infinity in spite of delays. Our results show that for smooth optimization problems, the delay is asymptotically negligible. In other words, the algorithm preserves the performance benefits despite using the stale gradient information available. We provide a provable guarantee that the explicit error bounds between the objective values of the estimates at each agent and the optimal value of the problem can quickly achieve a preset accuracy, when the stepsize is properly chosen.

*Notations*: The inner product of two vectors $x$ and $y$ is $\langle x, y\rangle$, and the standard Euclidean norm is defined by $\|x\| := \langle x, y\rangle^{1/2}$. The dual norm $\|\cdot\|_*$ to the Euclidean norm $\|\cdot\|$ is defined by $\|z\|_* := \sup_{\|x\| \le 1}\langle z, x\rangle$. Let $[A]_{ij}$ represent the $(i,j)$th element of a matrix $A$. For two matrices $A$ and $B$, we use $A - B \succcurlyeq 0$ to represent the matrix $A-B$ is positive semi-definite. For any $\epsilon \in [0,1]$, we say that a quantity $u$ is of order $\mathcal{O}(v(\epsilon))$ if there exists a constant $c>0$ such that $u \le cv(\epsilon)$.

## 2. Problem and algorithm

### 2.1. Problem formulation

Consider a time-varying network of $m$ agents, indexed by $i = 1, \ldots, m$. The communication topology is modeled by a directed graph $\mathcal{G}_t = (\mathcal{V}, \mathcal{E}_t)$ over the vertex set $\mathcal{V}$ with edge set $E_t \subset \mathcal{V} \times \mathcal{V}$, where $\mathcal{V} = \{1, \ldots, m\}$. Let $\mathcal{N}_i(t)$ represent the collection of in-neighbors, i.e., the set of agents that can send information to agent $i$ at time $t$. The network objective is to solve the following constrained optimization problem:

$$\min f(x) := \frac{1}{m}\sum_{i=1}^{m} f_i(x) \quad \text{s.t. } x \in X := \bigcap_{i=1}^{m} X_i, \tag{1}$$

where $f_i : \mathbb{R}^d \to \mathbb{R}$ is a local objective function at agent $i$, only known by agent $i$, and $X_i \subset \mathbb{R}^d$ is a local constraint at agent $i$. The intersection, $X$, of the constraint sets is assumed to be nonempty. For the simplicity of notation, we define $f^* = \min_{x \in X} f(x)$, $X^* = \{x \in X \mid f(x) = f^*\}$.

Our goal in this paper is to deal with the situation, in which each agent has only access to its private cost function $f_i$, $i \in \mathcal{V}$ and can communicate the information with its immediate neighbors.

### 2.2. Definitions and assumptions

In our algorithm, we make use of the following definitions.

**Definition 1** (*Bertsekas and Tsitsiklis [2]*). A differentiable function $\mu : \mathbb{R}^d \to \mathbb{R}$ is $c$-strongly convex, if there exists a constant $c>0$ such that for all $x, y \in \mathbb{R}^d$,

$$\mu(x) \ge \mu(y) + \langle \nabla\mu(y), x-y\rangle + \frac{c}{2}\|x-y\|^2,$$

where $\nabla$ is the gradient.

**Definition 2** (*Bregman [47]*). For a given strongly convex differentiable function $\mu : \mathbb{R}^d \to \mathbb{R}$, the Bregman divergence $D_\mu(x, y) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ between $x$ and $y$ based on a distance generating function $\mu$ is defined as follows:

$$D_\mu(x,y) = \mu(x) - \mu(y) - \langle \nabla\mu(y), x-y\rangle.$$

By Definitions 1 and 2, the Bregman divergence satisfies $D_\mu(x,y) \ge (c/2)\|x-y\|^2$. The Bregman divergence can be viewed as a natural generalization of the classical Euclidean squared