Automatica 83 (2017) 162–169

Contents lists available at ScienceDirect

Automatica

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

Brief paper

Convergence rate analysis of distributed optimization with projected subgradient algorithm[☆]

Shuai Liu [a,1], Zhirong Qiu [b,1], Lihua Xie [b]

[a] School of Control Science and Engineering, Shandong University, Jinan, 250061, China
[b] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, 639798, Singapore

ARTICLE INFO

Article history:
Received 23 June 2016
Received in revised form 2 May 2017
Accepted 4 May 2017

Keywords:
Distributed optimization
Subgradients
Convergence rate
Multi-agent systems

ABSTRACT

In this paper, we revisit the consensus-based projected subgradient algorithm proposed for a common set constraint. We show that the commonly adopted non-summable and square-summable diminishing step sizes of subgradients can be relaxed to be only non-summable, if the constrained optimum set is bounded. More importantly, for a strongly convex aggregate cost with different types of step sizes, we provide a systematical analysis to derive the asymptotic upper bound of convergence rates in terms of the optimum residual, and select the best step sizes accordingly. Our result shows that a convergence rate of $O(1/\sqrt{k})$ can be achieved with a step size $O(1/k)$.

1. Introduction

Distributed convex optimization over multi-agent networks has been receiving more and more research interest in the recent decade. One interesting and important case is that each agent has its own convex cost and decision variable, and needs to minimize the aggregate cost with an agreed decision variable, as in big data analysis (Cevher, Becker, & Schmidt, 2014) and distributed learning (Sayed, 2014). A variety of distributed optimization algorithms have been proposed, e.g. Chen and Ozdaglar (2012); Duchi, Agarwal, and Wainwright (2012); Iutzeler, Bianchi, Ciblat, and Hachem (2016); Nedić and Ozdaglar (2009); Nedić, Ozdaglar, and Parrilo (2010); Varagnolo, Zanella, Cenedese, Pillonetto, and Schenato (2016); Zhu and Martinez (2012), to name a few.

It is noted that many algorithms are (sub)gradients-based due to the inexpensive computation cost of (sub)gradients. Early works combined a weighted averaging of local decision variables with local (sub)gradient descent, and the convergence analysis explicitly relies on an asymptotic agreement of the decision variables. The distributed subgradient algorithm was firstly proposed in Nedić and Ozdaglar (2009), and later extended to the projected subgradient algorithm (Nedić et al., 2010) for a common set constraint. The case of nonidentical set constraints was studied in Lin, Ren, and Farrell (2016). By using push-sum techniques or surplus states, the above two algorithms can be further applied in directed networks with row stochastic weight matrices (Nedić & Olshevsky, 2015; Xi & Khan, 2016b). Distributed dual averaging methods were studied in Duchi et al. (2012) and Tsianos, Lawlor, and Rabbat (2012). By employing a Nestorov-type acceleration step, the fast gradient method and its proximal-gradient version are respectively studied in Jakovetic, Xavier, and Moura (2014) and Chen and Ozdaglar (2012) for multi-agent systems. In the above works, an exact convergence to the optimum requires either diminishing step sizes of (sub)gradients (Lin et al., 2016; Nedić & Olshevsky, 2015; Nedić et al., 2010; Xi & Khan, 2016b), or a drastically increasing communication burden of multi-step averaging during each iteration (Chen & Ozdaglar, 2012; Jakovetic et al., 2014). Both cases lead to a compromised convergence rate when compared with the corresponding centralized algorithms for differentiable functions. This issue has been addressed in recent works by including an additional averaging of local gradients, where a constant step size of gradients can be adopted to achieve a similar convergence rate as centralized algorithms (Nedić, Olshevsky, & Shi, 2016; Nedić, Olshevsky, Shi, & Uribe, 2016; Qu & Li, 2016; Shi, Ling, Wu, & Yin, 2015; Xi & Khan, 2016a, 2017; Xu, Zhu, Soh, & Xie, 2015).

Convergence rate is an important criterion for evaluating the performance of different optimization algorithms. Distributed subgradient methods can achieve a convergence rate of $O(\log k/\sqrt{k})$ for general convex costs (Nedić & Olshevsky, 2015; Xi & Khan,

[☆] This research is funded by Qilu Youth Scholar Discipline Construction Funding from Shandong University and National Natural Science Foundation of China (NSFC) under Grants 61304045, 61573220, 61633014. The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Michael M. Zavlanos under the direction of Editor Christos G. Cassandras.

E-mail addresses: LIUS0025@ntu.edu.sg (S. Liu), qiuz0005@e.ntu.edu.sg (Z. Qiu), elhxie@ntu.edu.sg (L. Xie).

[1] Both authors contributed equally to this paper and share the first authorship.

http://dx.doi.org/10.1016/j.automatica.2017.06.011
0005-1098/© 2017 Elsevier Ltd. All rights reserved.

2016b), and $O(\log k/k)$ for strongly convex functions (Nedić & Olshevsky, 2016). Distributed averaging methods achieve a convergence rate of $O(1/\sqrt{k})$ for general convex costs (Duchi et al., 2012; Tsianos et al., 2012). Note that these results are given as the ergodic rate in terms of the optimum value residual. When the cost function assumes Lipschitz continuous gradients, a sublinear rate of $O(1/k)$ can be achieved by adopting a constant step size (Qu & Li, 2016; Shi et al., 2015), and the fast distributed gradient method achieves a rate of $O(1/k^{2-\varepsilon})$ with an arbitrarily small ε (Jakovetic et al., 2014). If the cost function further assumes strong convexity, then a linear rate can be achieved (Nedić, Olshevsky, & Shi, 2016; Nedić, Olshevsky, Shi, & Uribe, 2016; Qu & Li, 2016; Shi et al., 2015; Xi & Khan, 2016a, 2017). Note that the rate analysis in above works is carried out for particular step sizes.

In this paper, we revisit the projected subgradient algorithm proposed in Nedić et al. (2010) under a common set constraint. We relax the non-summable and square-summable conditions of the diminishing step sizes of subgradients to non-summable when the constrained optimum set is bounded. Moreover, under the strong convexity condition we provide a systematic convergence rate analysis for different types of diminishing step sizes, and select the best step size accordingly. One specific choice of the best step size is given by $\alpha(k) = c/k$ where c is lower bounded by some constant, and the corresponding convergence rate is $O(1/\sqrt{k})$ in terms of optimum residual. When compared with the existing works which address non-smooth costs by subgradient methods, it outperforms the ergodic rate of $O(\log k/\sqrt{k})$ in Nedić and Olshevsky (2015) and Xi and Khan (2016b) with the additional strong convexity condition; besides, it also provides a sharper upper bound than the result in Nedić and Olshevsky (2016) where the convergence rate of optimum residual is given by $O(\sqrt{\frac{\log k}{k}})$.

The rest of the paper is organized as follows. Some preliminaries on graph theory and convex analysis are briefly reviewed in Section 2, and the problem formulation is introduced in Section 3 together with some necessary assumptions. The convergence under a relaxed condition of diminishing step sizes is proved in Section 4, followed by the corresponding rate analysis in Section 5. We conclude our work in Section 7.

Throughout this paper, the following notations are used. $\mathbb{R}$, $\mathbb{R}^n$ and $\mathbb{R}^{m\times n}$ represent the set of real numbers, the set of n-dimensional real column vectors and the set of $m \times n$ real matrices, respectively. $\mathbb{Z}^+$ stands for the set of positive integers. Given a matrix M, we use $[M]_{ij}$ to denote its (i,j)-th entry and M' its transpose. $\mathrm{col}\{x_1, \dots, x_n\}$ is a column vector with the ith block equal to x_i. $\mathbf{1}_n$ is an n-dimensional column vector with all elements being 1 and $J = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n'$. For $x, y \in \mathbb{R}^m$, $\langle x, y\rangle$ and $\|x - y\|$ are their inner product and the induced distance, i.e. $\langle x, y\rangle = x'y$, $\|x - y\| = \sqrt{\langle x-y, x-y\rangle}$. $\Psi_M(\cdot,\cdot)$ denotes the matrix multiplication deductively defined as $\Psi_M(i,j) = M(j)\Psi_M(i,j-1)$, $\forall i \le j$ and $\Psi_M(i,j) = I$, if $i > j$. Given two functions $\alpha_1(k), \alpha_2(k) : \mathbb{Z}^+ \to \mathbb{R}$, $\alpha_1(k) = O(\alpha_2(k))$ if $\limsup_{k\to+\infty}\left|\frac{\alpha_1(k)}{\alpha_2(k)}\right| < +\infty$ and $\alpha_1(k) = o(\alpha_2(k))$ if $\lim_{k\to+\infty}\frac{\alpha_1(k)}{\alpha_2(k)} = 0$.

2. Preliminaries

2.1. Graphs and nonnegative matrices

A multi-agent system and the communication among different agents (nodes) can be modeled as a directed graph $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$. $\mathcal{N} = \{1, \dots, n\}$ is the node set, and $\mathcal{E} = \{(i,j) : i, j \in \mathcal{N}\}$ is the edge set of ordered pairs, where $(i,j) \in \mathcal{E}$ implies that node j can receive information from node i. A nonnegative weight matrix $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ can be further assigned to a graph $\mathcal{G}$, where $a_{ij} > 0$ iff $(j,i) \in \mathcal{E}$ and $\sum_j a_{ij} = 1$ for each i. A path from node i to j is defined by an edge sequence $(i, n_1), (n_1, n_2), \cdots, (n_i, j) \in \mathcal{E}$. $\mathcal{G}$ is said to be strongly connected if a path can always be found between any two different nodes. $\mathcal{G}(k) = \{\mathcal{N}, \mathcal{E}(k)\}$ denotes the graph at each time instant k, and the joint graph over time span $[k_1, k_2]$ is given by $\mathcal{G}[k_1, k_2] = \{\mathcal{N}, \cup_{k=k_1}^{k=k_2}\mathcal{E}(k)\}$. We say that graphs $\{\mathcal{G}(k), k \in \mathbb{Z}^+\}$ are uniformly strongly connected, if there exists $\kappa \in \mathbb{Z}^+$ such that $\mathcal{G}[k, k+\kappa]$ is strongly connected for each k.

2.2. Convex analysis

2.2.1. Convex sets

A set $\mathcal{C} \subseteq \mathbb{R}^m$ is convex if $\theta x + (1-\theta)y \in \mathcal{C}$ for any $x, y \in \mathcal{C}$ and $\theta \in (0,1)$. For a closed convex set $\mathcal{C}$, $P_{\mathcal{C}}(\cdot) : \mathbb{R}^m \to \mathcal{C}$ is a projection operator which maps $x \in \mathbb{R}^m$ to a unique point $P_{\mathcal{C}}(x)$ such that $\|x - P_{\mathcal{C}}(x)\| = \inf_{v\in\mathcal{C}}\|x - v\| \triangleq \|x\|_{\mathcal{C}}$. $P_{\mathcal{C}}$ is non-expansive in the sense that

$$\|P_{\mathcal{C}}(x) - P_{\mathcal{C}}(y)\| \le \|x - y\|, \ \forall x, y \in \mathbb{R}^m. \tag{1}$$

Moreover, for any $y \in \mathcal{C}$, it holds that (Nedić & Ozdaglar, 2009)

$$\|P_{\mathcal{C}}(x) - y\|^2 \le \|x - y\|^2 - \|x\|_{\mathcal{C}}^2. \tag{2}$$

2.2.2. Convex functions

A real function f defined on $\mathbb{R}^m$ is convex if for any $x, y \in \mathbb{R}^m$ and $\theta \in (0,1)$, it holds that $f(\theta x + (1-\theta)y) \le \theta f(x) + (1-\theta)f(y)$. The subdifferential of a convex function f at x is defined by the set

$$\partial f(x) = \{g_f(x) : f(y) \ge f(x) + \langle g_f(x), y - x\rangle, \ \forall y\}, \tag{3}$$

and $g_f(x) \in \partial f(x)$ is called a subgradient of f at x.

Given a convex set $\mathcal{C}$, a convex function f is β-strongly convex over $\mathcal{C}$ if there exists $\beta > 0$ such that $\forall x, y \in \mathcal{C}, \theta \in (0,1)$,

$$f(\theta x + (1-\theta)y) \le \theta f(x) + (1-\theta)f(y) - \frac{\beta}{2}\theta(1-\theta)\|x - y\|^2. \tag{4}$$

Furthermore, if f attains the minimum over $\mathcal{C}$ at $x^* \in \mathcal{C}$, then f can be lower bounded by a quadratic function as follows (Hiriart-Urruty & Lemaréchal, 2001):

$$f(y) - f(x^*) \ge \frac{\beta}{2}\|y - x^*\|^2, \ \forall y \in \mathcal{X}. \tag{5}$$

3. Problem formulation

Consider the following constrained optimization problem of minimizing a sum of convex cost functions as

$$\min_{x\in\mathcal{X}} F(x) \triangleq \sum_{i=1}^{n} f_i(x), \tag{6}$$

where each f_i is the individual cost of agent i, and $\mathcal{X} \subseteq \mathbb{R}^m$ is a closed and convex constraint set. Denote the state of agent i as x_i, then (6) is equivalent to

$$\min_{x_i\in\mathcal{X}} \sum_{i=1}^{n} f_i(x_i), \ \text{s.t.}\ x_1 = \cdots = x_n. \tag{6'}$$

More precisely, if we denote

$$\mathcal{X}^* = \arg\min_{x\in\mathcal{X}} F(x), \tag{7}$$

then (7) is solved in terms of following conditions:

$$\lim_{k\to\infty} \|x_i(k) - x_j(k)\| = 0, \ \forall i, j \in \mathcal{N}, \tag{8a}$$

$$\lim_{k\to\infty} \|x_i(k)\|_{\mathcal{X}^*} = 0, \ \forall i \in \mathcal{N}. \tag{8b}$$