Kui Zhu; Yichen Zhang; Yutao Tang
Distributed optimization with inexact oracle

DISTRIBUTED OPTIMIZATION WITH INEXACT ORACLE

Kui Zhu, Yichen Zhang, and Yutao Tang

In this paper, we study the distributed optimization problem using approximate first-order information. We suppose the agent can repeatedly call an inexact first-order oracle of each individual objective function and exchange information with its time-varying neighbors. We revisit the distributed subgradient method in this circumstance and show its suboptimality under square summable but not summable step sizes. We also present several conditions on the inexactness of the local oracles to ensure an exact convergence of the iterative sequences towards the global optimal solution. A numerical example is given to verify the efficiency of our algorithm.

1. INTRODUCTION

Consider a network of m agents and each agent i corresponds to a convex objective function $f_i(x)$. Suppose the agent can collect the data of each individual function and exchange some information with other agents. We aim at an optimal solution and the optimal value of the optimization problem:

$$\min_{x \in \mathbb{R}^n} f(x) = \sum_{i=1}^{m} f_i(x) \tag{1}$$

This problem has been coined as distributed optimization and shown to have many applications in multi-robot systems, smart grids, communication networks, and large-scale machine learning [17, 31].

Among different classes of numerical algorithms for this problem, methods that use the first-order information (i. e., the function value $f_i(s)$ and its subdifferential $\partial f_i(s)$ at specific point $s \in \mathbb{R}^n$) have been intensively studied due to the cheap iteration cost and well-established convergence properties. In the seminal work [19], a subgradient rule was proposed to solve the distributed optimization problem, where each agent performs a consensus step and then a descending step along the subgradient direction of each local objective function. Since then, many attempts have been made to solve such distributed

optimization problems in different settings by primal domain, dual domain, or primal-dual methods, just to list a few [5, 8, 11–13, 25, 26, 30, 32, 33]. With the help of these computational results, recent extensions have been developed to incorporate non-trivial agent dynamics in such distributed optimization problems [27–29]. Note that these optimization results can be roughly regarded as distributed extensions of the classical centralized first-order methods and all require the exact first-order information of the individual objective functions.

In practice, the first-order information of the objective functions (especially the subgradients) may not be exactly available without knowing the function structure. It is usually calculated by considering another auxiliary optimization problem, which is often solved with certain accuracy [4, 6]. This issue will be aggravated by possible environment noises [16]. Thus it is crucial to consider the solvability of the given optimization problem using only approximate first-order information. In fact, this issue has been paid much attention to in the field of mathematical programming. For example, the well-known notion of ε-subgradient has been delivered and intensively used when the exact subgradient is not available [4, 16, 21, 23]. Some attempts have also been made to deal with the inexactness of gradients in distributed settings, e. g., [3, 15, 34, 35].

Recently, the authors in [7] introduced a very general notion of inexact first-order oracle to describe the inexactness of the first-order information for given objective functions. It has been intensively utilized to analyze the behavior of several typical centralized inexact first-order methods in the past few years and also been applied to study distributed optimization algorithms [9, 18, 22]. Typically, these papers are often focused on the evolution of some averaging dynamics while the oracle inexactness is interpreted from the coordinate error among agents to prove the convergence performance of the proposed distributed algorithms therein. Note that they still assume the exact first-order information of each individual objective function. With the aforementioned observation, a natural question arises: *Whether and how can we develop distributed first-order optimization methods with local inexact oracles to solve the optimization problem* (1)?

Having this question in mind, we will focus the optimization problem (1) and seek effective distributed algorithms to solve it providing that only approximate first-order information of the objective functions is available. To this end, we revisit the well-known distributed subgradient method proposed in [19] but working with an inexact oracle of each f_i. We first show the convergence of this rule under diminishing step sizes and then investigate its optimality or suboptimality under different oracle conditions.

The main contribution can be summarized as follows.

- We present a novel distributed algorithm to solve the consensus optimization problem under time-varying communication graphs. This rule can be taken a distributed counterpart of existing primal gradient method with inexact oracle in [7]. Compared with existing distributed algorithms in [8, 19, 25, 26, 30, 32], this rule can work with only approximate first-order information of the objective functions.

- We establish the convergence of our proposed algorithm under diminishing step sizes and show that it can approximately solve the problem provided that each local objective function has an inexact first-order oracle. Moreover, we specify a group of conditions under which the exact convergence of such iterative sequences towards the global optimal solution can be ensured even with inexact local oracles.

The rest of this paper is organized as follows. In Section 2, we first introduce some notations and concepts for the following analysis. Then, we present a novel distributed rule with inexact oracle in Section 3. After that, we consider the convergence and suboptimality issues of our algorithm under different conditions in Section 4. We also give a simulation example in Section 5 to illustrate the effectiveness of our design. Finally, we conclude the paper with some remarks in Section 6.

2. PRELIMINARY

We briefly present some preliminaries about our notations and terminology for the following analysis [4, 7].

2.1. Notation

Let $\mathbb{R}^n$ be the n-dimensional Euclidean space and $\mathbb{R}^{n \times m}$ be the set of all $n \times m$ matrices with all entries in $\mathbb{R}$. A vector is viewed as a column, unless clearly stated otherwise. $\mathbf{1}_n$ (or $\mathbf{0}_n$) denotes an n-dimensional all-one (or all-zero) column vector and $\mathbf{1}_{n \times m}$ (or $\mathbf{0}_{n \times m}$) all-one (or all-zero) matrix. $\mathrm{diag}(b_1, \ldots, b_n)$ denotes an $n \times n$ diagonal matrix with diagonal elements $b_1, \ldots, b_n$. $\mathrm{col}(a_1, \ldots, a_n) = [a_1^\top, \ldots, a_n^\top]^\top$ for column vectors a_i $(i = 1, \ldots, n)$. For a vector x (or matrix A) , $||x||$ $(||A||)$ denotes its Euclidean (or spectral) norm. When A is square, its trace and Frobenius norm are denoted by $\mathrm{Tr}(A)$ and $||A||_{\mathrm{F}} = \sqrt{\mathrm{Tr}(A^\top A)}$, respectively.

A column (or row) vector x is said to be a stochastic vector when its components are nonnegative and their sum is equal to 1. A square matrix is doubly stochastic when its rows are stochastic vectors, and its columns are also stochastic vectors.

2.2. Convex function and inexact oracle

Consider a function $f \colon \mathbb{R}^m \to \mathbb{R}$. It is said to be convex if for $0 \leq a \leq 1$ and $\forall \zeta_1, \zeta_2 \in \mathbb{R}^m$, we have

$$f(a\zeta_1 + (1 - a)\zeta_2) \leq af(\zeta_1) + (1 - a)f(\zeta_2).$$

If it is differentiable, the convexity of this function is equivalent with the following inequality:

$$f(\zeta_1) \geq f(\zeta_2) + \nabla f(\zeta_2)^T(\zeta_1 - \zeta_2), \forall \zeta_1, \zeta_2 \in \mathbb{R}^m.$$

Function $g \colon \mathbb{R}^m \to \mathbb{R}^m$ is Lipschitz with constant $\vartheta_1 > 0$ if

$$||g(\zeta_1) - g(\zeta_2)|| \leq \vartheta_1 ||\zeta_1 - \zeta_2||, \forall \zeta_1, \zeta_2 \in \mathbb{R}^m.$$

A convex function f is said to admit a first-order (δ, L)-oracle if for any given point $x \in \mathbb{R}^m$, we have a pair $(f_{\delta, L}(x), g_{\delta, L}(x)) \in \mathbb{R} \times \mathbb{R}^m$ such that

$$0 \leq f(y) - f_{\delta, L}(x) - g_{\delta, L}(x)^\top(y - x) \leq \frac{L}{2}||y - x||^2 + \delta$$

holds for any $y \in \mathbb{R}^m$ with $\delta \geq 0$ being the accuracy of the inexact first-order oracle. If f admits a (δ, L)-oracle, then cf admits a $(c\delta, cL)$-oracle for any constant $c > 0$. If f_i admits a (δ_i, L_i)-oracle, $i = 1, \ldots, m$, then $\sum_{i=1}^m f_i$ admits a $(\sum_{i=1}^m \delta_i, \sum_{i=1}^m L_i)$-oracle.

When a differentiable convex function f has L-Lipschitz gradients, it admits a $(0, L)$-oracle with $f_{0,L}(x) = f(x)$ and $g_{0,L}(x) = \nabla f(x)$. This is the exact first-order information of function f. When a (nonsmooth) convex function f can be well approximated by a smooth convex function $\overline{f}$ with Lipschitz gradients in the sense that their difference is bounded, the exact first-order information of function $\overline{f}$ provides an inexact oracle for function f.

3. DISTRIBUTED RULE WITH INEXACT ORACLE

The goal of this multi-agent system is to collectively minimize the global objective function f in a distributed manner. As usual, we assume that the optimal value $f^* = \min_{x \in \mathbb{R}^n} f(x)$ is finite and the optimal solution set $\mathcal{X}^* = \{x \in \mathbb{R}^n \mid f(x) = f^*\}$ is nonempty.

In this paper, we are more interested in the solvability of the distributed optimization problem working with only an inexact oracle of the local objective function.

Assumption 1. For each $i = 1, \ldots, m$, function f_i is convex and admits a first-order (δ, L)-oracle.

Several nontrivial classes of examples satisfying this assumption can be found in [7]. Although smooth convex functions with Lipschitz gradients naturally satisfy this assumption, f_i needs not to be differentiable at all. We will give a nonsmooth example later in Section 5.

We suppose agent i maintains an estimate $x_i(k)$ of the global optimal solution to problem (1) at time instant $k = 1, 2, \ldots$ and can share its estimate with other agents. Meanwhile, it will collect the data of problem (1) by calling the first-order (δ, L)-oracle of f_i at time k. Suppose the oracle will return a pair of $(f_{i,\delta}(k), g_{i,\delta}(k))$ satisfying the following assumption:

Assumption 2. There exists a constant $C > 0$ such that $\max_i\{\|g_{i,\delta}(k)\|\} \leq C$ for all $k \geq 0$ and $i = 1, \ldots, m$.

Each agent will update its estimate based on the inexact first-order information of the local objective function and all available estimates including its own and those received from others. To describe the communication topology, we introduce a sequence of undirected graphs $\{\mathcal{G}_k\}$ with node set $\mathcal{N} = \{1, \ldots, m\}$ and edge set $\mathcal{E}_k$. An edge $(i, j) \in \mathcal{E}_k$ means that agent i and agent j can communicate with each other at time instant k. Motivated by the well-known distributed subgradient method in [19], we consider the following iteration with inexact oracles of the individual functions:

$$x_i(k+1) = \sum_{j=1}^{m} w_{ij}(k)x_j(k) - \alpha_k g_{i,\delta}(k) \tag{2}$$

where $g_{i,\delta}(k)$ is an inexact subgradient of function f_i at point $x_i(k) \in \mathbb{R}$ by calling its inexact oracle and $\alpha_k > 0$ is the step size to be specified later. Here $w_{ij}(k)$ is a nonnegative weight with the property that $w_{ij}(k) > 0$ if and only if node i receives information from node j at time k. We use a matrix $W(k) = [w_{ij}(k)]_{m \times m}$ to represent these weights

and denote $\mathbf{x}(k) = \mathrm{col}(x_1(k), \ldots, x_m(k))$, $g_\delta(\mathbf{x}(k)) = \mathrm{col}(g_{1,\delta}(k), \ldots, g_{m,\delta}(k))$ for short. Then we can rewrite the distributed rule (2) into a compact form as follows:

$$\mathbf{x}(k+1) = [W(k) \otimes I_n]\mathbf{x}(k) - \alpha_k g_\delta(\mathbf{x}(k)). \tag{3}$$

We may denote $g_\delta(k) = g_\delta(\mathbf{x}(k))$ for short. This rule (2) extends the distributed subgradient method developed in [19] to the case when only approximate first-order information of the local objective function is available.

Here are some assumptions on the communication graphs to guarantee a consensus when $g_\delta(\mathbf{x}(k)) = \mathbf{0}$, which have been widely used in literature [14, 19, 20].

Assumption 3. Graph $\mathcal{G}_k$ is connected for any $k \geq 1$.

Assumption 4. For any $k \geq 1$, the matrix $W(k)$ is doubly stochastic. Additionally, there is a constant $\eta > 0$ such that, for any $i = 1, \ldots, m$, $w_{ii}(k) \geq \eta$ and $w_{ij}(k) \geq \eta$ if $(j, i) \in \mathcal{E}_k$.

In the following section, we will study the convergence of our algorithm (2) and show the (approximate) solvability of distributed optimization problem (1) by this rule with the inexact oracle. For the ease of presentation, we assume $n = 1$ and multiple dimensional analysis can be done for each component without any technical obstacles.

4. CONVERGENCE ANALYSIS

This section is dedicated to the performance analysis of the distributed rule (2). Denote the average of all agents' estimates at time k by $x_{\mathrm{av}}(k) = \frac{\mathbf{1}^\top \mathbf{x}(k)}{m}$ and let $\overline{\mathbf{x}}(k) = \mathbf{x}(k) - \mathbf{1}x_{\mathrm{av}}(k)$. It follows then

$$\overline{\mathbf{x}}(k+1) = Q(k)\overline{\mathbf{x}}(k) - \alpha_k R g_\delta(k) \tag{4}$$

where $Q(k) = W(k) - \frac{\mathbf{1}\mathbf{1}^\top}{m}$ and $R = I_m - \frac{\mathbf{1}\mathbf{1}^\top}{m}$.

We denote $\Phi(k, s) \triangleq \prod_{k'=s}^{k} W(k')$. The following lemma ensures an asymptotic consensus of our algorithm (2) under diminishing step sizes.

Lemma 1. Suppose Assumptions 1–4 hold and choose the step size $\{\alpha_k\}$ such that $\lim_{k\to\infty} \alpha_k = 0$. Along the trajectory of algorithm (2), all estimates of agents can achieve a consensus on their time-varying average, that is, $\overline{\mathbf{x}}(k) \to \mathbf{0}$ as $k \to \infty$.

P r o o f. To prove this lemma, we view it as a perturbed stability problem of (4) with a perturbation $\alpha_k R g_\delta(k)$. Note that $\prod_{k'=s}^{k} Q(k') = \Phi(k, s) - \frac{\mathbf{1}\mathbf{1}^\top}{m}$. It follows then

$$\overline{\mathbf{x}}(k+1) = \prod_{k'=K}^{k} Q(k')\overline{\mathbf{x}}(K) - \sum_{k'=K}^{k-1} \alpha_{k'} \prod_{i=k'+1}^{k} Q(i) R g_\delta(k') - \alpha_k R g_\delta(k)$$

$$= [\Phi(k, K) - \frac{\mathbf{1}\mathbf{1}^\top}{m}]\overline{\mathbf{x}}(K) - \alpha_k R g_\delta(k) - \sum_{k'=K}^{k-1} \alpha_{k'}[\Phi(k, k'+1) - \frac{\mathbf{1}\mathbf{1}^\top}{m}]R g_\delta(k')$$

for any $k > K$. We will show that for any $\varepsilon > 0$, there exist a large enough K such that $||\bar{\mathbf{x}}(k+1)|| < \varepsilon$ holds for any $k > K$.

Under the theorem conditions, we first recall Lemma 1 in [17]. It follows that

$$([\Phi(k, s)]_{ij} - \frac{1}{m})^2 \leq (1 - \frac{\eta}{2m^2})^{k-s}$$

holds for any $s > 0$ and $k \geq s$. This jointly with the fact $\sqrt{1-\mu} \leq 1 - \frac{\mu}{2}$ when $0 < \mu < 1$ implies that

$$|[\Phi(k, s)]_{ij} - \frac{1}{m}| \leq \sqrt{(1 - \frac{\eta}{2m^2})^{k-s}} \leq (1 - \frac{\eta}{4m^2})^{k-s}.$$

Denote $q = 1 - \frac{\eta}{4m^2}$ for short. Since $||[\Phi(k, s) - \frac{\mathbf{1}\mathbf{1}^\top}{m}]||_F \leq m \max_{i,j}\{|[\Phi(k, s)]_{ij} - \frac{1}{m}|\} \leq mq^{k-s}$. Noticing that $||g_\delta(k)|| \leq \sqrt{m}C$ under Assumption 2, one can obtain

$$||\bar{\mathbf{x}}(k+1)|| \leq mq^{k-K}||\bar{\mathbf{x}}(K)|| + \alpha_k \sqrt{m}C||R|| + m^{3/2}C||R|| \sum_{k'=K}^{k-1} \alpha_{k'} q^{k-k'-1}$$

$$\leq mq^{k-K}||\bar{\mathbf{x}}(K)|| + \sup_{k' \geq K}\{\alpha_{k'}\} \frac{m^{3/2}C||R||}{q(1-q)}.$$

Since $\lim_{k \to \infty} \alpha_k = 0$, for a given $\varepsilon > 0$, there exists a large enough K' such that $\sup_{k' \geq K'}\{\alpha_{k'}\} \leq \frac{q(1-q)\varepsilon}{2m^{3/2}C||R||}$. Note that there exists an integer $K > K'$ such that $q^{k-K'} < \frac{\varepsilon}{2m||\bar{\mathbf{x}}(K')||}$ when $k > K$ due to the fact that $0 < q < 1$. Therefore, we have $||\bar{\mathbf{x}}(k+1)|| < \varepsilon$ for any $k > K$. The proof is thus complete. □

Lemma 1 states that all agents' estimates of the global optimal solution eventually reach a consensus if $\lim_{k \to \infty} \alpha_k = 0$. To establish the convergence of (2), we shall focus on the evolution of x_{av} governed by

$$x_{\mathrm{av}}(k+1) = x_{\mathrm{av}}(k) - \alpha_k \frac{\mathbf{1}^\top g_\delta(k)}{m}. \tag{5}$$

Denote $\Delta_i(k) = g_{i,\delta}(k)(x_i(k) - x_{\mathrm{av}}(k))$. By the definition of (δ, L)-oracle, we have

$$f_i(y_i) - f_{i,\delta}(k) - g_{i,\delta}(k)(y_i - x_{\mathrm{av}}(k)) + \Delta_i(k) \geq 0$$
$$f_i(y_i) - f_{i,\delta}(k) - g_{i,\delta}(k)(y_i - x_{\mathrm{av}}(k)) + \Delta_i(k)$$
$$\leq L||y_i - x_{\mathrm{av}}(k)||^2 + L||x_{\mathrm{av}}(k) - x_i(k)||^2 + \delta.$$

Summing up these inequalities together, we have

$$\Xi(\mathbf{y}, k) \geq 0$$
$$\Xi(\mathbf{y}, k) \leq L||\mathbf{y} - \mathbf{1}x_{\mathrm{av}}(k)||^2 + L||\bar{\mathbf{x}}(k)||^2 + m\delta \tag{6}$$

where $\mathbf{y} = \mathrm{col}(y_1 \ldots, y_m)$, $\Xi(\mathbf{y}, k) \triangleq \sum_{i=1}^m f_i(y_i) - (\sum_{i=1}^m f_{i,\delta}(k) - \sum_{i=1}^m \Delta_i(k)) - \sum_{i=1}^m g_{i,\delta}(k)(y_i - x_{\mathrm{av}}(k))$.

It is interesting to remark that the above two inequalities can be interpreted in the terminology of inexact oracle. In fact, by taking $\mathbf{y} = \mathbf{1}y$ for any $y \in \mathbb{R}$, we can find that $\sum_{i=1}^{m} f_{i,\delta}(k) - \sum_{i=1}^{m} \Delta_i(k)$ and $\sum_{i=1}^{m} g_{i,\delta}(k)$ are the approximate first-order information of function f at the point $x_{\mathrm{av}}(k)$ with time-varying oracle accuracy $L\|\bar{\mathbf{x}}(k)\|^2 + m\delta$. Since $\|\bar{\mathbf{x}}(k)\|$ will eventually converge to zero under a diminishing step size, we are going to get an inexact oracle of f with accuracy $m\delta$ in an asymptotic sense. This fact has been extensively exploited in [9, 18, 22] to establish the convergence of the distributed optimization algorithms therein when $\delta = 0$.

With these observations, it is time to present the first main theorem of the paper.

Theorem 2. Suppose Assumptions 1–4 hold and choose a positive step size $\{\alpha_k\}$ to satisfy the condition

$$\sum_{k=1}^{\infty} \alpha_k = \infty, \quad \sum_{k=1}^{\infty} \alpha_k^2 < \infty.$$

Then, the following inequality holds for sequences $\{x_i(k)\}$ generated by algorithm (2):

$$\liminf_{k \to \infty} f(x_{\mathrm{av}}(k)) \leq f^* + m\delta.$$

P r o o f . To prove this theorem, we consider (5). For any $x^* \in \mathcal{X}^*$, we denote $\bar{x}(k) = x_{\mathrm{av}}(k) - x^*$ and obtain that

$$|\bar{x}(k+1)|^2 = |\bar{x}(k) - \alpha_k \frac{\mathbf{1}^\top g_\delta(k)}{m}|^2$$

$$= |\bar{x}(k)|^2 - 2\alpha_k \frac{\mathbf{1}^\top g_\delta(k)}{m}\bar{x}(k) + |\alpha_k \frac{\mathbf{1}^\top g_\delta(k)}{m}|^2.$$

By evaluating (6) at $\mathbf{y} = \mathbf{1}x^*$ and $\mathbf{y} = \mathbf{1}x_{\mathrm{av}}(k)$ respectively, the following inequalities hold:

$$\mathbf{1}^T g_\delta(k)\bar{x}(k) \geq -f(x^*) + (\sum_{i=1}^{m} f_{i,\delta}(k) - \sum_{i=1}^{m} \Delta_i(k))$$

$$f(x_{\mathrm{av}}(k)) - (\sum_{i=1}^{m} f_{i,\delta}(k) - \sum_{i=1}^{m} \Delta_i(k)) \leq L\|\bar{\mathbf{x}}(k)\|^2 + m\delta.$$

Then, under Assumption 2, we have

$$|\bar{x}(k+1)|^2 \leq |\bar{x}(k)|^2 + C^2\alpha_k^2 + \frac{2\alpha_k}{m}[f(x^*) - f(x_{\mathrm{av}}(k)) + L\|\bar{\mathbf{x}}(k)\|^2 + m\delta]$$

$$\leq |\bar{x}(k)|^2 + C^2\alpha_k^2 - \frac{2\alpha_k}{m}[-L\|\bar{\mathbf{x}}(k)\|^2 + f(x_{\mathrm{av}}(k)) - m\delta - f(x^*)].$$

We claim that $\liminf_{k \to \infty}(f(x_{\mathrm{av}}(k)) - m\delta - f(x^*)) \leq 0$ and try to seek a contradiction. For the purpose, we denote $b_0 = \liminf_{k \to \infty}(f(x_{\mathrm{av}}(k)) - m\delta - f(x^*))$ and assume $b_0 > 0$.

Since $\|\overline{\mathbf{x}}(k)\|$ goes to zero when $k \to \infty$, there must be a large integer $K > 0$ such that $f(x_{\mathrm{av}}(k)) - m\delta - f(x^*) - L\|\overline{\mathbf{x}}(k)\|^2 > \frac{b_0}{2}$ for any $k > K$. This implies that

$$|\overline{x}(k+1)|^2 \le |\overline{x}(k)|^2 + C^2\alpha_k^2 - \frac{\alpha_k}{m}b_0.$$

Or equivalently,

$$\frac{b_0}{m} \sum_{k'=K}^{k} \alpha_k' \le |\overline{x}(K)| + C^2 \sum_{k'=K}^{k} \alpha_{k'}^2$$

for any $k > K$, which is impossible for a sufficient large k under theorem conditions. Hence, we have found a contradiction. It is safe for us to conclude that $b_0 \le 0$. The proof is thus complete. $\qquad\square$

Remark 3. From Theorem 2, we conclude that the distributed optimization problem (1) is approximately solved by rule (2) in the sense that $\lim_{k\to\infty}[x_i(k) - x_{\mathrm{av}}(k)] = 0$ and $\liminf_{k\to\infty} f(x_{\mathrm{av}}(k)) \le f^* + m\delta$. On the one hand, this quantified relationship is consistent with the centralized results in [7, 16]. On the other hand, when the first-order oracle is exact with $\delta = 0$, we recover the convergence and optimality conclusions obtained in existing literature [9, 18, 22, 30].

Remark 4. It is remarkable that the oracle inexactness does not accumulate over time according to the rule (2). In fact, it can be verified that only the long-term accuracy matters in the steady-state performance of our algorithm if the oracle inexactness is time-varying. This property might be favorable in real-time noisy circumstances.

Next, we will specify some conditions on the oracle accuracy under which the exact convergence of sequences $\{x_i(k)\}$ to an optimal solution is guaranteed. Since diminishing oracle accuracy is required to ensure the exact (consensus) solvability of our problem by Theorem 2, we make another assumption to replace Assumption 1.

Assumption 5. For each $i = 1, \ldots, m$, function f_i is continuous and admits a first-order (δ_k, L_k)-oracle and $\max_k\{L_k\} \le L$ for some constant $L > 0$.

Here is the second main theorem of this paper.

Theorem 5. Suppose Assumptions 2–5 hold and the positive step size satisfies

$$\sum_{k=1}^{\infty} \alpha_k = \infty, \quad \sum_{k=1}^{\infty} \alpha_k^2 < \infty, \quad \sum_{k=1}^{\infty} \alpha_k\delta_k < \infty.$$

Then the following assert will take place:

a) The sequence $\{\|x_{\mathrm{av}}(k) - x^*\|\}$ converges;

b) There exists a cluster point $\overline{x}_{\mathrm{av}}$ of sequence $\{x_{\mathrm{av}}(k)\}$ such that $\overline{x}_{\mathrm{av}} \in X^*$;

c) If the optimal solution to problem (1) is unique, then $\lim_{k\to\infty} x_i(k) = x^*$ for each $i = 1, \ldots, m$.

P r o o f. Following the proof of Theorem 2, we have

$$|\overline{x}(k+1)|^2 \leq |\overline{x}(k)|^2 + C^2\alpha_k^2 - \frac{2\alpha_k}{m}[-L\|\overline{\mathbf{x}}(k)\|^2 + f(x_{\mathrm{av}}(k)) - m\delta_k - f(x^*)] \quad (7)$$

with $\overline{x}(k) = x_{\mathrm{av}}(k) - x^*$. Ignoring the negative terms and using the fact $m \geq 1$, we have

$$|\overline{x}(k+1)|^2 \leq |\overline{x}(k)|^2 + C^2\alpha_k^2 + 2L\alpha_k\|\overline{\mathbf{x}}(k)\|^2 + 2\alpha_k\delta_k.$$

We claim that $\sum_{k=1}^{\infty} \alpha_k\|\overline{\mathbf{x}}(k)\|^2 < \infty$.

We first show $\sum_{k=1}^{\infty} \alpha_k\|\overline{\mathbf{x}}(k)\| < \infty$. From the proof of Lemma 1, the following inequality holds for any $k \geq 3$:

$$\alpha_k\|\overline{\mathbf{x}}(k)\| \leq m\alpha_k q^{k-2}\|\overline{\mathbf{x}}(1)\| + \alpha_k\alpha_{k-1}\sqrt{m}C\|R\| + m^{3/2}C\|R\|\alpha_k \sum_{k'=1}^{k-2} \alpha_{k'} q^{k-k'-2}.$$

Completing the square to handle the α_k-related terms, one can further obtain that

$$\alpha_k\|\overline{\mathbf{x}}(k)\| \leq m^2\|\overline{\mathbf{x}}(1)\|^2\alpha_k^2 + q^{2(k-2)} + mC\|R\|(\alpha_k^2 + \alpha_{k-1}^2)$$
$$+ m^2C\|R\|\alpha_k^2 \sum_{k'=1}^{k-2} q^{k-k'-2} + m^2C\|R\| \sum_{k'=1}^{k-2} \alpha_{k'}^2 q^{k-k'-2}$$

where we use the fact $\sqrt{m} \leq m$ for $m \geq 1$. Summing up all from $k = 3$ and rearranging some terms, we have

$$\sum_{k=3}^{\infty} \alpha_k\|\overline{\mathbf{x}}(k)\| \leq (m^2\|\overline{\mathbf{x}}(1)\|^2 + 2mC\|R\| + \frac{m^2C\|R\|}{1-q}) \sum_{k=1}^{\infty} \alpha_k^2$$
$$+ \frac{1}{1-q^2} + m^2C\|R\| \sum_{k=1}^{\infty} \sum_{k'=1}^{k} \alpha_{k'}^2 q^{k-k'}.$$

Since $0 < q < 1$ and $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$, it must hold that $\sum_{k=1}^{\infty} \sum_{k'=1}^{k} \alpha_{k'}^2 q^{k-k'} < \infty$ by Lemma 7 in [20]. Thus, $\sum_{k=1}^{\infty} \alpha_k\|\overline{\mathbf{x}}(k)\| < \infty$. Recalling the fact that $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$ implies $\lim_{k\to\infty} \alpha_k = 0$, one can conclude that $\overline{\mathbf{x}}(k)$ will converge to 0 by Lemma 1. This jointly with $\sum_{k=1}^{\infty} \alpha_k\|\overline{\mathbf{x}}(k)\| < \infty$ means that the preceding claim is indeed true.

At this stage, we can directly use Lemma 2.1 in [2] to conclude the convergence of $|\overline{x}(k)|^2$ (i.e., item a)) and $\sum_{k=1}^{\infty} \alpha_k[f(x_{\mathrm{av}}(k)) - f^*] < \infty$ under the theorem conditions.

Since $|x_{\mathrm{av}}(k+1) - x_{\mathrm{av}}(k)| = |\alpha_k\frac{\mathbf{1}^\top g_\delta(k)}{m}| \leq C\alpha_k$ under Assumption 5, we resort to Proposition 2 in [1] and have $\lim_{k\to\infty} f(x_{\mathrm{av}}(k)) = f^*$. At the same time, the sequence $\{x_{\mathrm{av}}(k)\}$ is bounded by item a) and must have a convergent subsequence $\{\overline{x}_{k_{k'}}\}$. We denote the limit of this subsequence by $\overline{x}_{\mathrm{av}}$. Then, from the continuity of function f, we have $f(\overline{x}_{\mathrm{av}}) = f(\lim_{k'\to\infty} \overline{x}_{k_{k'}}) = \lim_{k'\to\infty} f(\overline{x}_{k_{k'}}) = f^*$, which means $\overline{x}_{\mathrm{av}} \in X^*$ (i.e., item b)).

When the optimal solution is unique, all convergent subsequences of $\{x_i(k)\}$ converge to the same limit. Then, $\lim_{k\to\infty} x_i(k) = \lim_{k\to\infty} x_{\mathrm{av}}(k) = x^*$ for each $i = 1, \ldots, m$. The proof is thus complete. $\square$
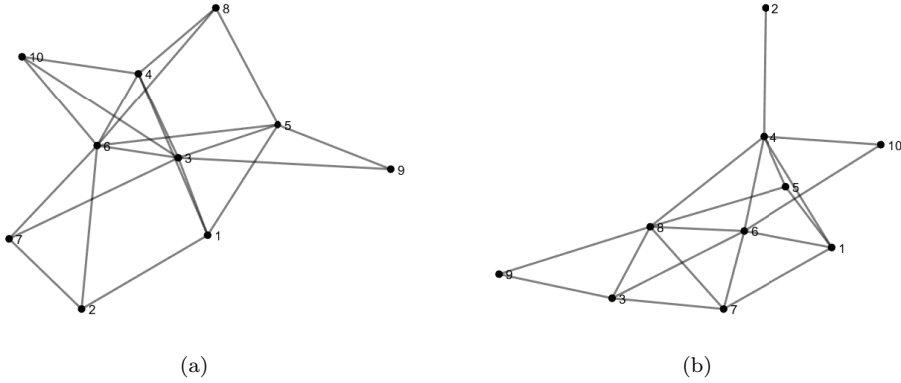
(a)   (b)

Fig. 1. Communication graphs in our example.

Remark 6. This theorem states that the distributed rule (2) can generate the exact optimal solution to problem (1) even with only some inexact first-order information provided that the accumulated oracle inexactness is finite in terms of $\sum_{k=1}^{\infty} \alpha_k \delta_k < \infty$. This condition is motivated by our previous work [35] on distributed primal-dual ε-subgradient method for such optimization problems. Similar conditions have been widely used to handle the inexactness of subgradient and proximal point calculation in centralized optimization scenarios [10, 16, 23, 24]. This property can be very useful especially when the first-order information is numerically obtained by solving another auxiliary optimization problem with controlled accuracy.

5. SIMULATION

In this section, we provide a numerical example to illustrate the effectiveness of our proposed algorithm in simulations.

Consider the following LASSO regression problem

$$\min_{x \in \mathbb{R}^n} \{\frac{1}{2}||Ax - y||^2 + \lambda ||x||_1\}$$

where $A \in \mathbb{R}^{N \times n}$, $y \in \mathbb{R}^N$, and $\lambda > 0$. Suppose the training data (y, A) is collected by m agents in the form: $A = [A_1, \ldots, A_m]^\top$ and $y = [y_1, \ldots, y_m]^\top$ with $y_i \in \mathbb{R}^{N_i}$, $A_i \in \mathbb{R}^{N_i \times n}$, and $\sum_{i=1}^m N_i = N$. The problem can be rewritten into the form (1) with $f_i(x) = \frac{1}{2}||A_i x - y_i||^2 + \frac{\lambda}{m}||x||_1$ ($i = 1, \ldots, m$) and solved by the distributed subgradient method [14, 19]. Here we use this example to show the effectiveness of our algorithm working with inexact first-order oracle.

Note that $|x|$ can be lower approximated by the following Huber loss function:

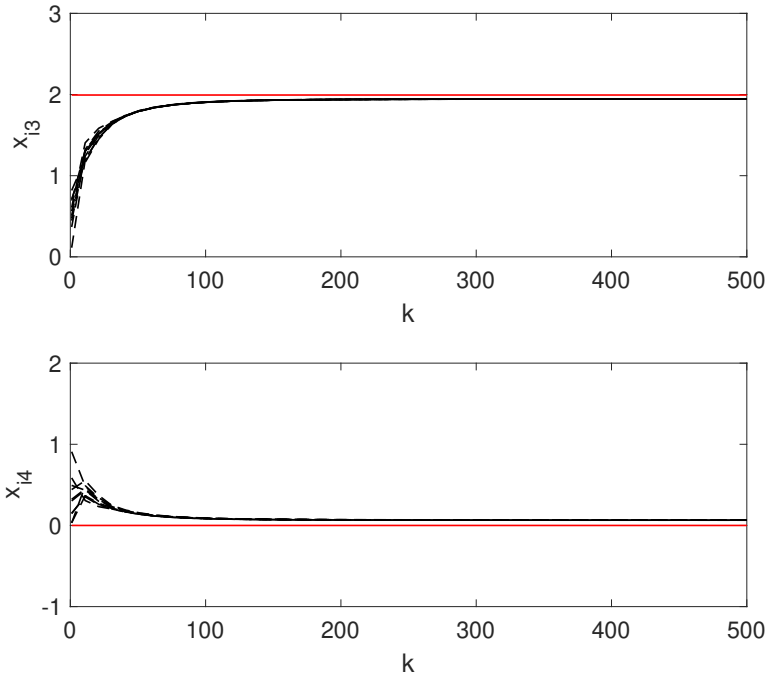$$H_\delta(x) = \begin{cases} x^2/(2\delta), & |x| \le \delta \\ |x| - \delta/2, & |x| > \delta. \end{cases}$$

Fig. 2. Profile of x_{i3} and x_{i4} with $\delta_k = 1$. The estimates are
illustrated by black dashed lines and the global optimal solution by
red solid lines.

We denote $\tilde{f}_{i\delta}(x_i) = \frac{1}{2}\|A_i x_i - y_i\|^2 + \frac{\lambda}{m}\sum_{i=1}^{n} H_\delta(x_{ij})$ for each $x_i = \mathrm{col}(x_{i1}, \ldots, x_{in}) \in$ $\mathbb{R}^n$ with $0 < \delta \le 1$. It can be verified that $\tilde{f}_{i\delta}(x_i)$ is convex, differentiable with Lipschitz gradient. Moreover, $(\tilde{f}_{i\delta}(x_i), \nabla\tilde{f}_{i\delta}(x_i))$ is an $(n\delta/2, \sqrt{n}/\delta + \|A_i\|^2)$-oracle for the nonsmooth objective function $f_i(x_i)$. Then we can apply the distributed rule (2) to resolving this LASSO regression problem.

In simulations, we let $N = 1000$, $m = 10$, $N_i = 100$, $n = 5$ and $\lambda = 1$. The matrix A is randomly chosen with each entry in $[0, 1]$ and $y = Ax_0 + \epsilon$ with $x_0 = \mathrm{col}(0, 1, 2, 0, 1)$ and ϵ randomly chosen in $[-0.01, 0.01]$. Suppose the communication topology alternates between the two graphs depicted in Figure 1 every 50 iterations. The corresponding two weighted matrices are randomly chosen to satisfy Assumption 4. The step size is chosen as $\alpha_k = 0.5/(k + 80)$ to ensure a consensus of all agents' estimates. For the inexact oracle, we first fix the oracle accuracy at $\delta_k = 1$ and then use the time-varying one with $\delta_k = 1/k$. The profiles of two components of agent i's estimate x_i (i.e., x_{i3} and x_{i4}) under different oracle conditions are reported in Figures 2 and 3. In Figure 2, with constant oracle accuracy, the estimates are observed to reach a consensus but deviate from the global optimal solution x^*. By contrast, these estimates converge to x^* when we use inexact oracle with time-varying accuracy δ_k as depicted in Figure 3. The profile of $\sum_{i=1}^{m} \tilde{f}_{i\delta}(x_i)$ in two cases and also the optimal value are shown in Figure 4. It can be
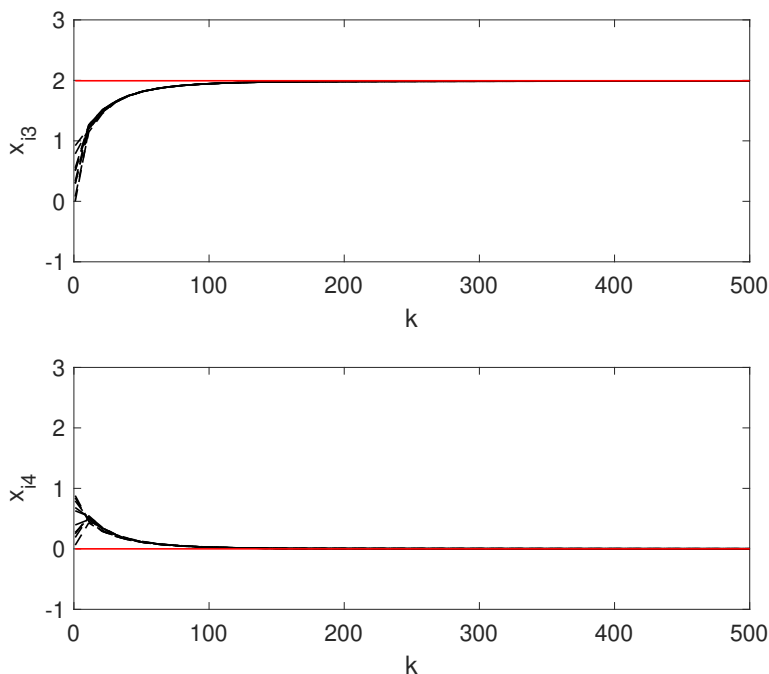
Fig. 3. Profile of x_{i3} and x_{i4} with $\delta_k = 1/k$. The estimates are illustrated by black dashed lines and the global optimal solution by red solid lines.

found that the distributed rule (2) can indeed produce an approximation of the optimal value when $\delta_k = 1$ while the term $\sum_{i=1}^{m} \tilde{f}_{i\delta}(x_i)$ eventually converges to f^* when $\delta_k = \frac{1}{k}$. These observations confirm the preceding theoretical conclusions on the proposed rule (2) to solve the distributed problem (1) with inexact first-order information.

6. CONCLUSION

This paper has attempted to solve a distributed optimization problem using only the approximate first-order information of the objective functions. We have proposed a variant of the distributed subgradient method working with inexact oracle and discussed its convergence and suboptimality properties under different stepsize and oracle conditions. In the future, we may consider the convergence rate of the iterative sequences and further exploit its deep connection with other classes of distributed optimization algorithms.
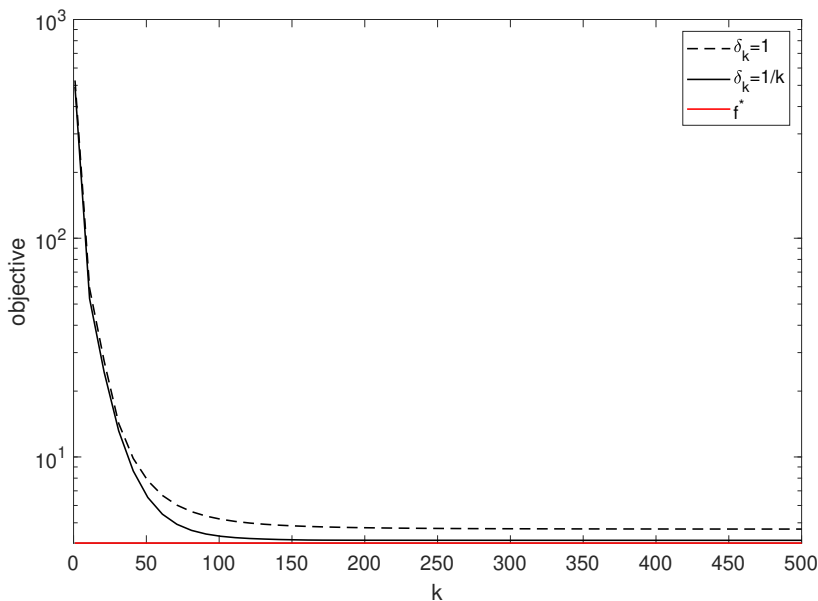
Fig. 4. Profile of $\sum_{i=1}^{m} \tilde{f}_{i\delta}(x_i)$ with different inexact oracles.

REFERENCES

[1] Y. I. Alber, A. N. Iusem, and M. V. Solodov: On the projected subgradient method for nonsmooth convex optimization in a Hilbert space. Math. Program. *81* (1998), 23–35. DOI:10.1007/bf01584842

[2] A. Auslender and M. Teboulle: Interior gradient and epsilon-subgradient descent methods for constrained convex minimization. Math. Oper. Res. *29* (2004), 1–26. DOI:10.1287/moor.1030.0062

[3] N. Bastianello and E. Dall'Anese: Distributed and inexact proximal gradient method for online convex optimization. In: Proc. European Control Conf. 2021, pp. 2432–2437. DOI:10.23919/ecc54610.2021.9654953

[4] D. P. Bertsekas: Convex optimization algorithms. Athena Scientific, Belmont 2015.

[5] S. Cheng and S. Liang: Distributed optimization for multi-agent system over unbalanced graphs with linear convergence rate. Kybernetika *56* (2020), 559–577. DOI:10.14736/kyb-2020-3-0559

[6] R. Correa and C. Lemaréchal: Convergence of some algorithms for convex minimization. Math. Program. *62* (1993), 261–275. DOI:10.1007/BF01585170

[7] O. Devolder, F. Glineur, and Y. Nesterov: First-order methods of smooth convex optimization with inexact oracle. Math. Program. *146* (2014), 37–75. DOI:10.1007/s10107-013-0677-5

[8] J. C. Duchi, A. Agarwal, and M. J. Wainwright: Dual averaging for distributed optimization: Convergence analysis and network scaling. IEEE Trans. Autom. Control *57* (2011), 592–606. DOI:10.1016/j.disamonth.2011.09.001

[9] D. Jakovetić, J. Xavier, and J. M. Moura: Fast distributed gradient methods. IEEE Trans. Automat. Control *59* (2014), 1131–1146. DOI:10.1109/TAC.2014.2298712

[10] K. C. Kiwiel: Convergence of approximate and incremental subgradient methods for convex optimization. SIAM J. Optim. *14* (2004), 807–840. DOI:10.1137/S1052623400376366

[11] P. Li and J. Hu: A solution strategy for distributed uncertain economic dispatch problems via scenario theory. Control Theory Technol. *19* (2021), 499–506. DOI:10.1007/s11768-021-00068-6

[12] P. Li, J. Hu, L. Qiu, Y. Zhao, and B. K. Ghosh: A distributed economic dispatch strategy for power-water networks. IEEE Trans. Control Netw. Syst. *9* (2022), 356–366. DOI:10.1109/TCNS.2021.3104103

[13] P. Li, Y. Zhao, J. Hu, Y. Zhang, and B. K. Ghosh: Distributed initialization-free algorithms for multi-agent optimization problems with coupled inequality constraints. Neurocomputing *407* (2020), 155–162. DOI:10.1016/j.neucom.2020.05.006

[14] S. Liu, Z. Qiu, and L. Xie: Convergence rate analysis of distributed optimization with projected subgradient algorithm. Automatica *83* (2017), 162–169. DOI:10.1016/j.automatica.2017.06.011

[15] Y. Lou, G. Shi, K. Johansson, and Y. Hong: Approximate projected consensus for convex intersection computation: Convergence analysis and critical error angle IEEE Trans. Automat. Control *59* (2014), 1722–1736. DOI:10.1109/TAC.2014.2309261

[16] A. Nedić and D. P. Bertsekas: The effect of deterministic noise in subgradient methods. Math. Program. *125* (2010), 75–99. DOI:10.1007/s10107-008-0262-5

[17] A. Nedić and J. Liu: Distributed optimization for control. Ann. Rev. Control Robot. Auton. Syst. *1* (2018), 77–103. DOI:10.1146/annurev-control-060117-105131

[18] A. Nedić, A. Olshevsky, and W. Shi: Achieving geometric convergence for distributed optimization over time-varying graphs. SIAM J. Optim. *27* (2017), 2597–2633. DOI:10.1137/16M1084316

[19] A. Nedic and A. Ozdaglar: Distributed subgradient methods for multi-agent optimization. IEEE Trans. Automat. Control *54* (2009), 48–61. DOI:10.1109/TAC.2008.2009515

[20] A. Nedić, A. Ozdaglar, and P. A. Parrilo: Constrained consensus and optimization in multi-agent networks. IEEE Trans. Automat. Control *55* (2010), 922–938. DOI:10.1109/TAC.2010.2041686

[21] B. Polyak: Introduction to optimization. Optimization Software Inc., New York 1987.

[22] G. Qu and N. Li: Harnessing smoothness to accelerate distributed optimization. IEEE Trans. Control Netw. Syst. *5* (2018), 1245–1260. DOI:10.1109/TCNS.2017.2698261

[23] J. Rasch and C. Antonin: Inexact first-order primal-dual algorithms. Comput. Optim. Appl. *76* (2020), 381–430. DOI:10.1007/s10589-020-00186-y

[24] R. T. Rockafellar: Monotone operators and the proximal point algorithm. SIAM J. Control Optim. *14* (1976), 877–898. DOI:10.1137/0314056

[25] G. Shi, K. H. Johansson, and Y. Hong: Reaching an optimal consensus: dynamical systems that compute intersections of convex sets. IEEE Trans. Automat. Control *58* (2013), 610–622. DOI:10.1109/TAC.2012.2215261

[26] W. Shi, Q. Ling, G. Wu, and W. Yin: EXTRA: An exact first-order algorithm for decentralized consensus optimization. SIAM J. Optim. *25* (2015), 944–966. DOI:10.1137/14096668X

[27] Y. Tang, Z. Deng, and Y. Hong: Optimal output consensus of high-order multia-
     gent systems with embedded technique. IEEE Trans. Cybernet. *49* (2019), 1768–1779.
     DOI:10.1109/TCYB.2018.2813431

[28] Y. Tang and X. Wang: Optimal output consensus for nonlinear multiagent systems
     with both static and dynamic uncertainties. IEEE Trans. Automat. Control *66* (2021),
     1733–1740. DOI:10.1109/TAC.2020.2996978

[29] Y. Tang and K. Zhu: Optimal consensus for uncertain multi-agent systems by output
     feedback. Int. J. Robust Nonlinear Control *32* (2022), 2084–2099. DOI:10.1002/rnc.5928

[30] C. Xi and U. A. Khan: Distributed subgradient projection algorithm over directed graphs.
     IEEE Trans. Automat. Control *62* (2017), 3986–3992. DOI:10.1109/TAC.2016.2615066

[31] T. Yang, X. Yi, J. Wu, Y. Yuan, D. Wu, Z. Meng, Y. Hong, H. Wang, Z. Lin, and K. H.
     Johansson: A survey of distributed optimization. Ann. Rev. Control *47* (2019), 278–305.
     DOI:10.1016/j.arcontrol.2019.05.006

[32] P. Yi, Y. Hong, and F. Liu: Distributed gradient algorithm for constrained optimization
     with application to load sharing in power systems. Syst. Control Lett. *83* (2015), 45–52.
     DOI:10.1016/j.sysconle.2015.06.006

[33] X. Zeng, P. Yi, and Y. Hong: Distributed continuous-time algorithm for constrained
     convex optimizations via nonsmooth analysis approach. IEEE Trans. Automat. Control
     62 (2017), 5227–5233. DOI:10.1109/TAC.2016.2628807

[34] Y. Zhang, Z. Deng, and Y. Hong: Distributed optimal coordination for mul-
     tiple heterogeneous EulerLagrangian systems.   Automatica *79* (2017), 207–213.
     DOI:10.1016/j.automatica.2017.01.004

[35] K. Zhu and Y. Tang: Primal-dual ε-subgradient method for distributed optimization. J.
     Syst. Sci. Complex. (2022), accepted.

*Kui Zhu, School of Artificial Intelligence, Beijing University of Posts and Telecommu-
nications, Beijing 100876. P. R. China.*
     e-mail: 1239366542@qq.com

*Yichen Zhang, School of Artificial Intelligence, Beijing University of Posts and Telecom-
munications, Beijing 100876. P. R. China.*
     e-mail: zhangyc930@163.com

*Yutao Tang, Corresponding author. School of Artificial Intelligence, Beijing University
of Posts and Telecommunications, Beijing 100876. P. R. China.*
     e-mail: yttang@bupt.edu.cn