KYBERNETIKA — VOLUME *52* (2016), NUMBER 6, PAGES 898–913

AVERAGING APPROACH TO DISTRIBUTED CONVEX OPTIMIZATION FOR CONTINUOUS-TIME MULTI-AGENT SYSTEMS

Wei Ni and Xiaoli Wang

Recently, distributed convex optimization has received much attention by many researchers. Current research on this problem mainly focuses on fixed network topologies, without enough attention to switching ones. This paper specially establishes a new technique called averaging-base approach to design a continuous-time distributed algorithm for convex optimization problem under switching topology. This idea of using averaging was proposed in our earlier works for the consensus problem of multi-agent systems under switching topology, and it is further developed in this paper to gain further insight into the distributed optimization algorithm. Key techniques are used, such as two-time-scale analysis and asymptotic expansions for the solutions of the backward equation or Liouvill equation. Important results are obtained, including weak convergence of our algorithm to the optimal solution.

Keywords: distributed convex optimization, averaging approach, two-time-scale, Markovian switching, invariant measure

Classification: 93C15, 93C35

1. INTRODUCTION

Distributed convex optimization problems appear as an important area that is receiving a lot of interest from a variety of research communities such as network control systems, traffic flow optimization, sensor fusion, and machine learning. Distributed convex optimization refers to minimizing the aggregate sum of N convex cost functions by designing N algorithms distributed on N nodes of a network, with each convex function being available at a node and with each states of neighbor node algorithms being exchanged across the network, achieving the goal of letting all of the states of the agents consensually and asymptotically converge to the minimizer. Aside from the particular node algorithms, the importance of the network structure should be highlighted as a facilitator of implementing the distributed optimization strategy. As can be seen, the more complex of the network structure, the more difficulty the distributed optimization problem becomes. Of particular importance and challenge is the time-varying network topology. Because of technique difficulties, distributed convex optimization under time-varying network is

DOI: 10.14736/kyb-2016-6-0898

investigated with relatively few attention. This constitutes the standing point of the present paper.

Let us elaborate on this with detailed review of literatures. Roughly speaking, distributed convex optimization algorithms can be divided into discrete-time and continuous-time. Most of the available algorithms, such as the widely used distributed algorithms based on subgradient [14] and projected subgradient [15], are developed in discrete-time mainly due to the overwhelming ability of digital computers to execute the algorithms discretely. Recently, more and more distributed convex optimization algorithms are explored in continuous-time since continuous-time set up is favored for utilizing more techniques (the elegant Lyapunov argument in [4] for example) to prove the algorithm convergence, and is beneficial for adopting differential geometry viewpoint which is extremely powerful when the optimization is constrained (see for example [21]).

Along the line of continuous-time, the works in [23] (see also [24]) is among the first to devoted to the distributed convex optimization algorithms in continuous-time (DCO-CT), without giving proofs of algorithm convergence. Then [6] presents a proof and analyzes the distributed continuous-time convex optimization in more detail by using tools from nonsmooth analysis and set-valued dynamical systems. This algorithm is also extended in [13] to include additive persistent noise, so that a stochastic distributed optimization algorithm on weight-balanced digraph is formed. These algorithms, including many others [10], are second order in nature since the graph Laplacian is used twice. Similar second order algorithm is also proposed in [22] which applies an observer to the distributed convex optimization problem. As for other forms of results, the work [12] indicates a zero-gradient-sum algorithm which evolve invariantly on a zero-gradient-sum manifold and converge asymptotically to the unknown optimizer. Also, the group led by Hong report on this problem many effective algorithms such as the approximate-projection-based DCO-CT algorithm[11], the distributed primal-dual continuous-time gradient algorithm [27], the initial-free DCO-CT algorithm [28], internal model based DCO-CT [25], potential game base algorithm [26], nonsmooth analysis based algorithm [29], just to mention a few.

Almost all the existing optimization results were obtained based on the assumption that the network is fixed. To better understand how the switches of the network topology affects the distributed optimal algorithms, this paper specially establishes a new technique called averaging-base approach to design a continuous-time distributed algorithm for convex optimization problem. The idea of using averaging was proposed in our earlier works [17, 18, 19] for the consensus problem of multi-agent systems under switching topology. As will be seen, the dynamics for the continuous-time distributed convex optimization is a switched system which is difficult to analyze. We will construct a time-invariant system (term as average system later) to approximate this switched system and let the convergence analysis and design of the distributed optimization be established based on the average system. Therefore, many design methods for the fixed topology case can be resorted to. This is the basic idea of averaging, with details being developed in what follows. Specifically, the averaging method was generalized from determined case in [17, 18] to stochastic case in [19], leading to the conclusion that the averaging system obtained in [19] is exactly the one whose infinitesimal generator is the average of the generators of each subsystem. This unified viewpoint is further explored in this paper by presenting theoretical insight and asymptotic expansion technique.

Although the idea is simple, the core of the averaging is to construct the average

system. Our method is build on the multiscale analysis, followed by the expansions of the solutions to the backward equation (or Liouvill equation) into a series in term of a small parameter α which characterizes the fast switching of the network. The existence of a fast process obviously accompanies with a slow one. Discussions are in order. First, due to the large dimension of the real word network, the adding or deleting of links would frequently occur. Therefore, the Markov process $\sigma(t)$ describing the time dependence of the network change faster than the node dynamical process $X(t)$. We call $\sigma(t)$ the fast process and $X(t)$ the slow process. To model this fast time-varying property of σ, we re-scale the time scale t of σ, by using a small positive number α, as t/α to obtain a fast time-varying process $\sigma(t/\alpha)$. Obviously, the smaller $\alpha > 0$ is, the faster the network switches. That is, the parameter α characterizes the speed of the fast switching. Needless to say, we are more interested in evolution of the slow process $X(t)$ than that of the fast process $\sigma(t/\alpha)$. Secondly, we perform a procedure to eliminate the fast process $\sigma(t/\alpha)$ from the dynamic of $(X(t), \sigma(t/\alpha))$. This is realized by working on the backward equation associated with our proposed switching dynamics for the optimization problem. It is well known that the backward equation includes almost all the useful properties of the switching optimization dynamics. However, its analytical solutions are hard to obtain. Instead, we expand this solution into a series in term of α. Then, by the aid of egodic theorem and Fredholm Alternative theorem, we prove the first term in the series is nothing but a solution to another backward equation, called by us the average backward equation. This average backward equation, taking the PDE form, easily induces a dynamics in ODE, called by us the average system. It is this average system that we seek to.

Now, the attention flows to the relationship between the average system and our proposed switched system for distributed optimization under switched network. Note that corresponding to above two systems, there are a backward equation and an average backward equation, with the former having a series solution and the latter having a first order approximation solution. Since the first term in the series is a first-order approximation, the average backward equation is viewed as the first-order approximation of the backward equation, and therefore, the average dynamics corresponding to the average backward equation is a first order approximation of our proposed switched dynamics for distributed optimization under switching network. It is our knowledge that there are no results reported on the applications of averaging theory to the distributed optimization problem. Although our ideas are based on the tradition averaging techniques such as in [16], their results are not directly applicable to our scenario of distributed convex optimization. The method presented in this paper can provide a deep understanding and a good guidance in the design and analysis of distributed convex optimization algorithms under switching topologies.

2. PROBLEM FORMULATION AND PRELIMINARIES

2.1. Problem formulation

Consider the following optimization problem

$$\text{minimize} \quad \tilde{f}(x) = \sum_{i=1}^{N} f_i(x), \tag{1}$$

where $f_i : \mathbb{R}^n \to \mathbb{R}$ is a convex function. That is, the problem is to find $x^* \in \mathbb{R}^n$ such that the objective function $\tilde{f}(x)$ in (1) is minimized. Such an x^* is called an optimal solution and the corresponding value $p^* = \tilde{f}(x^*)$ is called optimal value. Throughout this paper, we assume that the optimal value of this problem is finite and the optimal solution set $\{x \in \mathbb{R}^n | \sum_{i=1}^N f_i(x) = p^*\}$ is nonempty and compact. This problem finds many applications on network optimization. For example, it arises in optimization problems in wireless sensor networks with $f_i(x)$ corresponding to the data collected by the ith sensor in the network (see [20]), or in neural network training problems with $f_i(x)$ corresponding to the ith training data set (see [1]).

Optimization over network usually proceeds in a distributed way. More specially, related to the optimal problem (1), there are N agents with each standing for a dynamical system whose state $x_i \in \mathbb{R}^n, i \in \{1, \ldots, N\}$, is viewed as the estimate of the optimal solution x^*. These agents exchange their estimates with others and the information exchange among them forms a network. In this setting, the algorithm design should include the following two objectives: (a) design the node dynamics which can only get access to data $f_i(x_i)$ and states from its neighboring agents and (b) find conditions on the network structure. The design is to achieve the goal that these agents cooperatively and asymptotically compute the optimal solution x^* in the sense of

$$\lim_{t \to +\infty} \big(x_1(t), \ldots, x_N(t)\big) = (x^*, \ldots, x^*). \tag{2}$$

The above problem is called as distributed convex optimization problem. The node dynamics can be designed either in discrete-time or in continuous-time. Also, the network can be fixed or time varying. This paper focuses on distributed convex optimization in continuous-time under time varying network.

2.2. Network structure

The structure of information exchange among N agents $\{1, 2, \ldots, N\}$ is described by a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, 2, \ldots, N\}$ is the set of nodes representing N agents and $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is the set of edges of the graph. An edge of $\mathcal{G}$ is denoted by (i, j) representing that agents i and j can exchange information between them. The graph considered in this paper is undirected in the sense that the edges (i, j) and (j, i) in $\mathcal{V}$ are considered to be the same. Two nodes i and j are neighbors if $(i, j) \in \mathcal{V}$. The set of neighbors of node i is denoted by $\mathcal{N}_i = \{j \in \mathcal{V} : (j, i) \in \mathcal{E}, j \neq i\}$. A path is a sequence of connected edges in a graph. A graph is connected if there is a path between every pair of nodes. The union of a collection of graphs defined on a given node set is a new graph whose edge set is the union of the edge sets of the members. We use the symbol $\cup$ to denote the graph union. We say that such a collection is jointly connected if the union of its members is a connected graph.

To describe the structure of a graph, one usually uses matrices. Let a_{ij} be the weight of edge (i, j). Obviously, $a_{ij} = a_{ji}$. The adjacency matrix A of a graph $\mathcal{G}$ on vertex $\{1, \ldots, N\}$ is an $N \times N$ matrix, whose off-diagonal entry on the (ij) position is a_{ij} if (i, j) is an edge of $\mathcal{G}$ and 0 otherwise, and whose diagonal entry on the (i, i) position is $-\sum_{j \in \mathcal{N}_i} a_{ij}$. Obviously, for any undirected graph, its Laplacian is symmetric, positive semi-definite, and satisfies $L \cdot \mathbf{1} = 0 \cdot \mathbf{1}$, where $\mathbf{1}$ is a column vector whose entries

are all one and whose dimension is determined from context. Furthermore, the graph Laplacian has the following refined property: The weighted graph $\mathcal{G}$ is connected if and only if Laplacian L of $\mathcal{G}$ has a simple zero eigenvalue.

This paper will deal with time-varying graph. Suppose there are S possible graphs $\{\mathcal{G}_1, \ldots, \mathcal{G}_S\}$ and among them the network structure is switched according to switching law $\sigma : [0, \infty) \to \mathbb{S}$ which is a right continuous and piece-wise constant mapping, where $\mathbb{S} = \{1, 2, \ldots, S\}$. The time dependence of the network structure is denoted by $\mathcal{G}_{\sigma(t)}$, with its corresponding Laplacian being denoted by $L_{\sigma(t)}$. Also, the neighbor of each agent i is time varying and it is denoted by $\mathcal{N}_i(t)$.

2.3. Distributed optimization dynamics

To approach the convex optimization problem (1) is a distributed way, we let the node dynamics be specified as follows,

$$\dot{x}_i = \sum_{j \in \mathcal{N}_i(t)} (x_j - x_i) - s(t) \nabla f_i(x_i), i = 1, \ldots, N, \tag{3}$$

where $s(t) > 0$ satisfies

$$\lim_{t \to +\infty} s(t) = 0, \quad \int_0^{+\infty} s(t)\, \mathrm{d}t = +\infty, \tag{4}$$

which will be used later to prove the convergence of the algorithm (3). From (3) we see that agent i only involves the states of its neighboring agents as well as the value of f_i evaluated at its own state.

For ease of presentation, we define $X = (x_1^T, \ldots, x_N^T)$ and $F(X) = (f_1(x_1), \ldots, f_N(x_N))^T$, and write the dynamics (3) in a compact form as

$$\dot{X}(t) = -\mathscr{L}_{\sigma(t)} X(t) - s(t) \nabla F(X(t)), \tag{5}$$

where $\nabla F(X) = (\nabla^T f_1(x_1), \ldots, \nabla^T f_N(x_N))^T$ and $\mathscr{L}_{\sigma(t)} = L_{\sigma(t)} \otimes I_n$. Obviously, $X^* \triangleq (x^*, \ldots, x^*)$ is an equilibrium of system (5). If we can show that X^* is asymptotically stable, then the distributed algorithm (5) cooperatively and asymptotically compute the optimal solution x^*.

The time-varying network in this paper is described by a continuous-time Markov chain $\sigma(t) : [0, +\infty) \to \mathbb{S}$ adapted to the filtration $\{\mathcal{F}_t | t \geq 0\}$. The random switching of the network makes stability of (5) difficult to analyze. To overcome this difficulty, we will propose an average system, which is essentially time-invariant, to approximate the trajectory of (5) by adopting the idea of the averaging principle, which has been applied successfully to our earlier work ([17, 18, 19]) on multi-agent systems.

2.4. Preliminary lemmas

The first lemma on optimization condition can be found in most optimization textbook such as [2].

Lemma 2.1. (Optimization Condition) Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be convex and suppose that $x^\star \in \mathbb{R}^n$ is a point at which f is differentiable. Then $x^\star$ is a solution to the optimization problem minimize$_{x\in\mathbb{R}^n} f(x)$ if and only if $\nabla f(x^\star) = 0$.

The second lemma is a popular result in functional analysis (see for example [3, pp. 641 Theorem 5(iii)]), but we only need a simple version in finite dimension. The third is about cocoercivity of a function.

Lemma 2.2. (Fredholm Alternative) Consider the solvability of linear algebraic systems of the inhomogeneous form $Ax = b$ with $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$. Then this equation either has a unique solution for every b or the homogeneous equation $Ax = 0$ has nontrivial solutions. More precisely, the inhomogeneous equation is solvable if and only if b belongs to the column space of A, which is the orthogonal complement of $ker(A^T)$.

The third lemma is on the definition of cocoercive and its property which can be found in [6] or [7]. For $\mu > 0$, a locally Lipschitz function $F : \mathbb{R}^n \to \mathbb{R}$ is called μ-cocoercive if for $X_1, X_2 \in \mathbb{R}^n$, the inequality holds $(X_1 - X_2)^T[\nabla F(X_1) - \nabla F(X_2)] \geq \mu\|\nabla F(X_1) - \nabla F(X_2)\|^2$. The cocoercive of a function can be characterized by the following lemma.

Lemma 2.3. (Cocoercivity Characterization) Let F be a differential convex function. Then ∇F is globally Lipschitz with Lipschitz constant $k > 0$ if and only if F is $1/k$-cocoercive.

The fourth lemma is about weak convergence which can be found in [8]. Let $\mathcal{B}$ be a metric space and $\mathscr{B}$ be the Borel σ-algebra there. Let $\{\mathbb{P}_\alpha\}_{\alpha>0}$ be a family of probability measures on $(\mathcal{B}, \mathscr{B})$. We say that $\mathbb{P}^\alpha$ converges weakly to a probability measure $\mathbb{P}$ on $(\mathcal{B}, \mathscr{B})$ if $\int f \, d\mathbb{P}_\alpha \overset{\alpha\to 0}{\longrightarrow} \int f \, d\mathbb{P}$ for all continuous and bounded functions $f : \mathcal{B} \to \mathbb{R}$. A family X^α of stochastic variables on the space $\mathbb{B}$ convergence weakly to the random variable on $\mathbb{B}$ if the measures $\mathbb{P}_\alpha$ corresponding to X^α convergence weakly to the measure $\mathbb{P}$ corresponding to X. In this paper, the metric space $\mathcal{B}$ will be taken as $\mathcal{B} = \mathcal{C}([0, \infty))$ equipped with the topology of uniform convergence on compact subsets of $\mathcal{C}([0, \infty))$. With this topology, we get the corresponding Borel σ-algebra $\mathscr{B}$. The following lemma characterizes the weak convergence.

Lemma 2.4. (Weak Convergence) The random variables X^α convergence weakly to X as $\alpha \to 0$, if and only if $E_{\mathbb{P}_\alpha}[f(X^\alpha)] \to E_{\mathbb{P}}[f(X)]$ as $\alpha \to 0$ for each bounded and continuous function $f : \mathcal{B} \to \mathbb{R}$.

3. AVERAGING APPROACH TO DISTRIBUTED CONVEX OPTIMIZATION

We use a two-time-scale method to construct for switched system (5) an average equation, which is time-invariant and approximates (5) in a proper sense. The distributed convex optimization design can be based the the average system. Convergence results are presented.

3.1. Two time scale optimization dynamics

Although the joint process $(X(t), \sigma(t))$ in (5) is a Markov process, the interested process $X(t)$ alone is not. Thus the basic tool from Markov theory can not be used directly here. Obviously, it is the switching network or presence of $\sigma(t)$ that makes the analysis difficult. To overcome this difficulty, we will use an averaging procedure in this section to operate on the time-varying system (5) to average out the fast time-varying process $\sigma(t)$, yielding a time-invariant systems called average system, whose solution will approximate $X(t)$.

The above procedure makes use of the fast time-varying property of $\sigma(t)$. Due to the large dimension of the real word network, the adding or deleting of links would frequently occur. Therefore, the Markov process $\sigma(t)$ describing the time dependence of the network change faster than the node dynamical process $X(t)$. We call $\sigma(t)$ fast process and $X(t)$ slow process. To model this fast time-varying property of $\sigma(t)$, we rescale $\sigma(t)$ by using a small positive number α to obtain a fast time-varying process $\sigma(t/\alpha)$. Obviously, the smaller $\alpha > 0$ is, the faster the network switches. After the time rescale, we obtain the following two-time scale dynamics

$$\dot{X}^{\alpha}(t) = -\mathscr{L}_{\sigma(t/\alpha)} X^{\alpha}(t) - s(t)\nabla F(X^{\alpha}(t)). \tag{6}$$

The stochastic process $(X^{\alpha}(t), \sigma(t))$ determined by (6) is an Markovian process, whose infinitesimal is $\mathcal{A}^{\alpha} = \frac{1}{\alpha}\mathcal{Q} + \mathcal{A}_s$, where $\mathcal{A}_s$ is the infinitesimal of the s-subsystem of (5), $\mathcal{Q}$ is the infinitesimal of the Markov chain $\sigma(t)$.

In what follows, we will use α to present an averaging procedure to average out the fast switching process $\sigma(t/\alpha)$, yielding an average equation which is time-invariant. To this, some conditions on σ are firstly assumed.

3.2. Conditions on network structure

We assume that $\sigma(t)$ is a continuous-time Markov chain $\sigma(t)$ defined on the finite state $\mathbb{S}$. A continuous-time Markov chain will have samples that exhibit jumps from one state to another. The statistics of the Markov chain $\sigma(t)$ is characterized by an initial probability distribution $\pi_0 = [\pi_{01}, \ldots, \pi_{0S}]^T$ defined over $\mathbb{S} = \{1, 2, \ldots, S\}$ with $\pi_{0i} = \mathbb{P}(\sigma(0) = i)$, and by a Metzler matrix $\mathcal{Q} = (q_{ij})_{S\times S} \in \mathbb{R}^{S\times S}$, which is also called the infinitesimal generator of the Markov chain and it describes the transition probability as follows:

$$\mathbb{P}\{\sigma(t+\Delta t) = j | \sigma(t) = i\} = \begin{cases} q_{ij}\Delta t + o(\Delta t), & i \neq j, \\ 1 + q_{ii}\Delta t + o(\Delta t), & i = j, \end{cases}$$

where $\Delta t > 0$, and $q_{ij} \geq 0$, for $i \neq j$, is the transition rate from mode i at time t to mode j at time $t + \Delta t$, and $q_{ii} = -\sum_{j\neq i} q_{ij}$, $\lim_{\Delta t\to\infty} o(\Delta t)/\Delta t = 0$. At time t the state of the Markov chain is determined according to the probability distribution $\pi(t) = (\pi_1(t), \ldots, \pi_S(t))^T$ with $\pi_s(t)$ being the probability that at time t the Markov system is in state s. The normalization condition $\sum_{s=1}^{S} \pi_s(t) = 1$ is usually assumed. Letting $\Delta t \to 0$, the infinitesimal form of the Markov dynamics can be written as $\dot{\pi}_s(t) = \sum_{i=}^{S} \pi_i(t) q_{is}$, $s = 1, \ldots, S$. In a compact form, the probability distribution $\pi(t)$ of $\sigma(t)$ obeys the differential equation

$$\dot{\pi}(t) = \mathcal{Q}^T \pi(t). \tag{7}$$

Since the Markov chain $\sigma(t)$ used in this paper is a finite state, it follows from [5, pp. 150–151] that that $|q_{ij}| < \infty$.

Note that the finite state set $\mathbb{S}$ of the the Markov chain $\sigma(t)$ can be partitioned uniquely according to the decomposition theorem [5] as $\mathbb{S} = \{\mathbb{T}, \mathbb{C}_1, \ldots, \mathbb{C}_{S-1}\}$, where $\mathbb{T}$ is the set of transient states and $\mathbb{C}_1, \ldots, \mathbb{C}_{S-1}$ are irreducible closed sets of recurrent states. We assume that, for each $i = 1, \ldots, S-1$, the union graph $\cup_{j\in\mathbb{C}_i}\bar{\mathcal{G}}_j$ is connected. Since the $\sigma(t)$ will take value only in one of the closed set $\mathbb{C}_i$ after a certain time instant, and also since we are interested in the asymptotic behavior of the system, we will assume that probability distribution $\pi(t)$ of $\sigma(t)$ is stationary, and it is denoted by π, which is defined by

$$\mathcal{Q}^T\pi = 0, \quad \sum_{i=1}^{S}\pi_i = 1, \quad \pi_i > 0. \tag{8}$$

In conclusion, we make the following assumption.

Assumption 3.1. The switching topologies described by a finite state Markov process $\sigma(t)$ have a stationary probability distribution $\pi = (\pi_1, \ldots, \pi_N)^T$ verifying (8), and the union graph $\cup_{s\in\mathbb{S}}\bar{\mathcal{G}}_s$ is connected.

Remark 3.2. The set of equations in (8) can be viewed as a characterization of the ergodicity for σ, and it is equivalent to saying that the null space of the adjoint generator $\mathcal{Q}^T$ consists of only constants. Since the infinitesimal generator is an operator acting on vector functions defined on the state space $\mathbb{S}$ of σ, any function in the null space is independent of the state of $s \in \mathbb{S}$.

3.3. Averaging system

This subsection proposes an averaging procedure for (6) to construct an average ODE, which does not depend on the switching signal σ. This average ODE acts as the role of approximating the original equation (5) in the weak sense. The properties of the average system is explored in next subsection.

Let ϕ be a real-valued function defined on the state space (X^α, σ) of (6) which is chosen sufficiently smooth, and let $W(t, X, s) = \mathbb{E}\left[\phi(X^\alpha(t), \sigma(t))|X(0) = X, \sigma(0) = s\right]$, where the expectation is taken over the randomness caused by the initial states as well as the stochastic switching. From the standard analysis in stochastic theory, $W(t, X, s)$ is a unique bounded classical solution to the following partial differential equation with the initial data $W(0, X, s) = \phi(X, s)$,

$$\frac{\partial}{\partial t}W(t, X, s) = \frac{1}{\alpha}\mathcal{Q}\overrightarrow{W}(t, X)[s] + \mathcal{A}_s W(t, X, s), \tag{9}$$

where $\overrightarrow{W}(t, X) = \big(W(t, X, 1), \ldots, W(t, X, S)\big)^T$ and $\mathcal{Q}\overrightarrow{W}(t, X)[s]$ denotes the s-row of the matrix $\mathcal{Q}\overrightarrow{W}(t, X)$. The partial differential equation (9) is termed as the backward Kolmogorov equation associated with the stochastic differential equation (6). Although the solution of the backward equation (9) includes some fundamental properties of the the stochastic differential equation (6), the analytic form of this solution is hard to

obtain. To avoid this difficulty, we seek an approximate solution of the form $W = W_0 + \alpha W_1 + \mathcal{O}(\alpha^2)$. Inserting this expression into (9) and equating coefficients of α^{-1} on both sides yields $\mathcal{Q}\overrightarrow{W}_0 = 0$. By the Assumption 3.1 and Remark 3.2, this equation implies that $\overrightarrow{W}_0$ is a function independent of the switching mode $s \in \mathbb{S}$; that is, $W_0(t, X, 1) = W_0(t, X, 2) = \ldots = W_0(t, X, S)$. For ease of notation, we denote them by $W_0(t, X)$. We will show that W_0 is a solution to another backward equation (Liouville equation) associated with an ODE which will be specified later. Similarly, inserting the expression of W into (9) and equating coefficients of α^0 on both sides yields

$$\mathcal{Q}\overrightarrow{W}_1 = \begin{pmatrix} \frac{\partial W_0}{\partial t} - \mathcal{A}_1 W_0 \\ \vdots \\ \frac{\partial W_0}{\partial t} - \mathcal{A}_S W_0 \end{pmatrix}.$$

This equation, together with the Freddhom alterative in Lemma 2.2, tells us that $(\frac{\partial W_0}{\partial t} - \mathcal{A}_1 W_0, \ldots, \frac{\partial W_0}{\partial t} - \mathcal{A}_S W_0)^T$ is perpendicular to the null space of $\mathcal{Q}^T$. Noting that $\text{Null}(\mathcal{Q}^T) = \pi$ in (8), it is nature to have $\sum_{s=1}^{S} \pi_s \left(\frac{\partial W_0}{\partial t} - \mathcal{A}_s W_0\right) = 0$, which gives rise to an *average backward equation*

$$\frac{\partial W_0(t, X)}{\partial t} = \bar{\mathcal{A}}\, W_0(t, X), \tag{10}$$

where $\bar{\mathcal{A}} = \sum_{s=1}^{S} \pi_s \mathcal{A}_s$ is the stochastic average of the infinitesimal generators $\mathcal{A}_s$ with respect to the invariant measure π. Associated with this average backward equation (10), one can construct an ODE, called the average ODE whose infinitesimal generator is $\bar{\mathcal{A}}$. Indeed, it can be constructed as

$$\dot{X}(t) = -\bar{\mathscr{L}} X(t) - s(t)\nabla F(X(t)), \tag{11}$$

where $\bar{\mathscr{L}} = \sum_{s=1}^{S} \pi_s \mathscr{L}_s$ is the average Laplacian and it corresponds to a weighted graph formed by $\cup_{s=1}^{S} \pi_s \mathcal{G}_s$, here the symbol $\pi_s \mathcal{G}_s$ means a weighted graph forming by multiplying each weight of $\mathcal{G}_s$ by π_s. For the partial differential equation (10), chose the initial data to be the same as that of (9), that is, $W_0(0, X) = \phi(X, s)$, where (and also in (12)) s is considered as a constant. Then $W_0(t, X)$ can be written as

$$W_0(t, X) = \mathbb{E}\left[\phi((X(t), s) | X(0) = X)\right], \tag{12}$$

where $X(t)$ is the solution of the average equation (11) and the expectation is taken over the randomness caused only by the initial state since s is now a constant.

So far, we obtain two backward equations (9) and (10), as well as two corresponding differential equations (6) and (11). Note that $W(t, X, s)$ is the solution to (9), and the first term $W_0(t, X)$ in the series for $W(t, X, s)$ is the solution to (10). Since $W(t, X, s) \overset{\alpha \to 0}{\longrightarrow} W_0(t, X)$ for each t, that is,

$$\mathbb{E}\left[\phi(X^\alpha(t), \sigma(t)) | X(0) = X, \sigma(0) = s\right] \overset{\alpha \to 0}{\longrightarrow} \mathbb{E}\left[\phi((X(t), s) | X(0) = X)\right], \forall t,$$

then by Lemma 2.4, $X^\alpha(t)$ convergents weakly to $X(t)$ as $\alpha \to 0$. Therefore, the above analysis yields the following theorem.

Theorem 3.3. The trajectory $X^\alpha(t)$ of the distributed optimization algorithm (6), with the switching network satisfying Assumption 3.1 and $s(t)$ satisfying (4), converges weakly to the trajectory $X(t)$ of the average system (11) as $\alpha \to 0$.

3.4. Distributed optimization by the average system

We have obtained in Theorem 3.3 that the solution of the distributed optimization system (6) can be approximated by its average system (11). We now show for the average system (11) that: (a) The average system (11) can achieve consensus, and (b) The consensus state of the average system (11) is exactly $X^* = (x^*, \ldots, x^*)^T$.

We first address the first issue. It is justified by the following theorem which is not only interest in itself, but also used in subsequent development.

Theorem 3.4. Under the Assumption 3.1, the average system (11) can achieve consensus $\lim_{t\to\infty} \|x_i(t) - x_j(t)\| = 0$ for any $i, j \in \{1, \ldots, N\}$.

Proof. Define $J = \frac{1}{n}\mathbf{1}\mathbf{1}^T$ and $\tilde{X} = (I - J)X$. It is then obvious that $\tilde{X} = 0$ if and only if $x_1 = \ldots = x_N = \frac{1}{N}(x_1 + \ldots + x_N)$. Noting that $(I - J)^2 = I - J$, one has

$$\begin{aligned}\dot{\tilde{X}}(t) &= (I - J)[-\bar{\mathscr{L}}X(t) - s(t)\nabla F(X(t)] \\ &= -\bar{\mathscr{L}}(I - J)X(t) - s(t)(I - J)\nabla F(X(t)) \\ &= -\bar{\mathscr{L}}\tilde{X}(t) - s(t)(I - J)\nabla F(X(t)).\end{aligned}$$

It then follows from the formula of variation that

$$\tilde{X}(t) = e^{-\bar{\mathscr{L}}t}\tilde{X}(0) - \int_0^t e^{-\bar{\mathscr{L}}(t-\tau)}\delta(\tau)\,\mathrm{d}\tau,$$

where $\delta(t) = s(t)(I - J)\nabla F(X(t))$. Since $\lim_{t\to+\infty} s(t) = 0$ and ∇F is bounded, then $\lim_{t\to+\infty} \delta(t) = 0$. That is, for any $\varepsilon > 0$, there exists $t_\varepsilon > 0$ such that $\|\delta(t)\| < \varepsilon$ for all $t \geq t_\varepsilon$. Therefore,

$$\|\tilde{X}(t)\| \leq \|e^{-\bar{\mathscr{L}}t}\tilde{X}(0)\| + \left\|\int_0^{t_\varepsilon} e^{-\bar{\mathscr{L}}(t-\tau)}\delta(\tau)\,\mathrm{d}\tau\right\| + \left\|\int_{t_\varepsilon}^t e^{-\bar{\mathscr{L}}(t-\tau)}\delta(\tau)\,\mathrm{d}\tau\right\|. \tag{13}$$

The estimation of $\|\tilde{X}(t)\|$ requires the fact that $\|e^{-\bar{\mathscr{L}}t}\| \leq e^{-\bar{\lambda}_2 t}$ for any $t \geq 0$, where $\bar{\lambda}_2$ is the second smallest eigenvalue of the Laplacian for the weighted graph $\cup_{s=1}^S \pi_s \mathcal{G}_s$ and it is guaranteed to be positive by the joint connectivity of the graph $\cup_{s=1}^S \pi_s \mathcal{G}_s$ due to Assumption 3.1. Indeed, along the trajectories $y(t)$ of the system $\dot{y} = -\bar{\mathscr{L}}y$, the time derivative of $V_1(y) = y^T y$ satisfies $\frac{d}{dt}V_1(y(t)) = -2y^T\bar{\mathscr{L}}y \leq -2\bar{\lambda}_2 V_1(y)$, impling $V_1(y(t)) \leq e^{-2\bar{\lambda}_2 t}V_1(y(0))$ or $\|y(t)\| \leq e^{-\bar{\lambda}_2 t}\|y(0)\|$, which is further used to calculate the matrix norm $\|e^{-\bar{\mathscr{L}}t}\| = \sup\{\|e^{-\bar{\mathscr{L}}t}y(0)\|/\|y(0)\| : y(0) \neq 0\} = \sup\{\|y(t)\|/\|y(0)\| : y(0) \neq 0\} \leq e^{-\bar{\lambda}_2 t}$.

With above inequality and under the assumption on the boundness of ∇F, the $\tilde{X}(t)$ in (13) can be further estimated as

$$\begin{aligned}\|\tilde{X}(t)\| &\le e^{-\bar{\lambda}_2 t}\|\tilde{X}(0)\| + \int_0^{t_\varepsilon} e^{-\bar{\lambda}_2(t-\tau)}\|\delta(\tau)\|\,\mathrm{d}\tau + \int_{t_\varepsilon}^{t} e^{-\bar{\lambda}_2(t-\tau)}\|\delta(\tau)\|\,\mathrm{d}\tau \\ &\le e^{-\bar{\lambda}_2 t}\|\tilde{X}(0)\| + \int_0^{t_\varepsilon} e^{-\bar{\lambda}_2(t-\tau)}\|\delta(\tau)\|\,\mathrm{d}\tau + \varepsilon\int_{t_\varepsilon}^{t} e^{-\bar{\lambda}_2(t-\tau)}\,\mathrm{d}\tau \\ &\le e^{-\bar{\lambda}_2 t}\|\tilde{X}(0)\| + \int_0^{t_\varepsilon} e^{-\lambda_2(t-\tau)}\|\delta(\tau)\|\,\mathrm{d}\tau + \varepsilon/\bar{\lambda}_2.\end{aligned}$$

Taking limits on both side yields $\lim_{t\to+\infty}\|\tilde{X}(t)\| \le \varepsilon/\bar{\lambda}_2$. By the arbitrariness of ε, one has $\lim_{t\to+\infty}\|\tilde{X}(t)\| = 0$. That is, $\lim_{t\to\infty}\|x_i(t) - x_j(t)\| = 0$ for all $i, j \in \{1, 2, \ldots, N\}$. □

We have shown in Theorem 3.4 that the average system (11) converges to a consensus state. We proceed to show that the consensus state is a constant vector and this vector is nothing but the optimal solution X^*.

Theorem 3.5. Under Assumption 3.1, the average system (11) converges asymptotically to the optimal solution $X^* = (x^*, \ldots, x^*)^T$.

Proof. Let $V(X) = \|X - X^*\|^2$ which obviously satisfies $V(X) \ge 0$ and $V(X) = 0$ if and only if $X = X^*$. Then the time derivative of V along the solutions $X(t)$ of (11) can be calculated as

$$\begin{aligned}\dot{V}(t) &= 2[X(t) - X^*]^T[-\bar{\mathscr{L}}X(t) - s(t)\nabla F(X(t))] \\ &\le -2s(t)[X(t) - X^*]^T[\nabla F(X(t)) - \nabla F(X^*)] \\ &\le -2\mu s(t)\|\nabla F(X(t)) - \nabla F(X^*)\|^2 \\ &= -2\mu s(t)\|\nabla F(X(t))\|^2 \le 0,\end{aligned} \tag{14}$$

where the last inequality uses the definition of μ-cocoercivity for the function F that $(X - X^*)^T[\nabla F(X) - \nabla F(X^*)] \ge \mu\|\nabla F(X) - \nabla F(X^*)\|^2$. The above inequality tells us that the limit $\lim_{t\to+\infty} V(X(t))$ exists and so does $\lim_{t\to+\infty} X(t)$. Denote this limit by $X^{\#}$. Noting that we have proved in Theorem 3.4 that $\lim_{t\to+\infty} x_i = \lim_{t\to+\infty} x_j$, we conclude $X^{\#} = \mathbf{1}\otimes x^{\#}$ for some $x^{\#} \in \mathbb{R}^n$. That is $\lim_{t\to+\infty} X(t) = \mathbf{1}\otimes x^{\#}$, meaning that consensus state is a constant.

It now remains to prove that the consensus state is exactly the optimal solution. To this, we rewrite the inequality (14) into integral form,

$$V(t) \le V(0) - 2\mu\int_0^t s(\tau)\|\nabla F(X(\tau))\|^2\,\mathrm{d}\tau. \tag{15}$$

On the other hand, according to Lemma 2.3, we have $\|\nabla F(X) - \nabla F(X^*)\| \le \frac{1}{\mu}\|X - X^*\|$ or $V(X) \ge \mu^2\|\nabla F(X)\|^2$. This and (15) give

$$\mu^2\|\nabla F(X(t))\|^2 \le V(0) - 2\mu\int_0^t s(\tau)\|\nabla F(X(\tau))\|^2\,\mathrm{d}\tau.$$

Then by the Gronwall's inequality, we have

$$\|\nabla F(X(t))\|^2 \leq \frac{V(0)}{\mu^2} \exp\left(-\frac{2}{\mu}\int_0^t s(\tau)\,\mathrm{d}\tau\right).$$

Noting that $\int_0^{+\infty} s(t)\,\mathrm{d}t = +\infty$ and ∇F is continuous, and letting $t \to +\infty$, one has $\|\nabla F(X^{\#})\| = 0$ or $\nabla F(X^{\#}) = 0$. It then follows from Lemma 2.1 that $X^{\#} = X^*$; that is, $\lim_{t\to+\infty} X(t) = X^*$, which says that the average system (11) converges asymptotically to the optimal solution X^*. □

3.5. Results and concluding remark

For the optimization problem (1), we design a distributed gradient descent algorithm (3) or (5) under a time-varying network. In view of the fast switching property of the network, the time scale of the Markov chain describing the network topology is rescaled by a small parameter $\alpha > 0$. The distributed convex algorithm in (5) is then transformed under the rescaled time into a two-scale dynamics (6), composing a fast process and a slow process. The fast process is averaged out by the aid the averaging procedure, and an average equation which depends only on the slow process is constructed. This average equation is viewed as an approximation of (5) in the weak sense, presented in Theorem 3.3. Furthermore, the average system is shown to achieve consensus to the optimal solution, included in Theorems 3.4 and 3.5. Therefore, summarizing Theorems 3.3 – 3.5 gives rise to the following theorem whose proof is straightforward.

Theorem 3.6. Under Assumption 3.1, the solutions of the distributed optimization system (6) converge asymptotically and consensually to the optimal solution X^* in the weak sense.

Remark 3.7. The distributed convex optimization problem (1) is solved in existing literatures by transforming it into an equivalent form of minimizing $\sum_{i=1}^N f_i(x_i)$ subject to $\mathscr{L}X = 0$; however, the transformed and the original optimization problems are equivalent if and only if the graph is fixed and connected (see for example [6]). Therefore this routine is not applied in our paper since the graph may be disconnected at each time.

Remark 3.8. The algorithm for distributed convex optimization in continuous-time was firstly proposed in [23, 24], but they were second order. The work [6] also proposed a continuous-time algorithm, but of the observer type. The algorithm presented in this paper is first order and thus it is less complicated than them for implementation.

4. SIMULATION

Consider the optimization problem (1) with $N = 5$ and

$$f_1(x) = e^{x+2}, \quad f_2(x) = (4x+5)^2, \quad f_3(x) = (7x+17)^2$$
$$f_4(x) = (x+2)^4, \quad f_5(x) = 4.$$

The graph of the function $F(x) = \sum_{i=1}^{N} f_i(x)$ is plotted in Figure 1. The true optimal solution and optimal value of this problem are found, by using *fminbnd* in MatLab, to be $x^* = -2.1450$ and $F(x^*) = 21.6221$. We now used our distributed convex optimization algorithm to check the results.

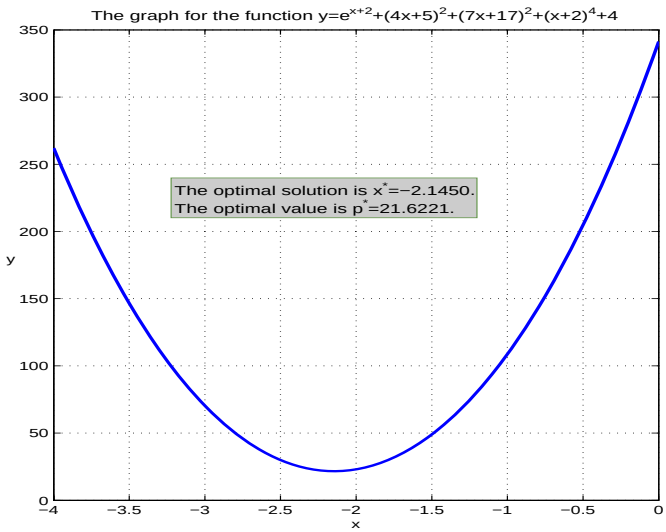

Fig. 1. The graph of the objective function.

To apply the distributed convex algorithm (3) to find the optimal solution and optimal value of this problem, our simulation uses a time-varying network which randomly switching among six possible undirected graphs $\{\mathcal{G}_1, \ldots, \mathcal{G}_6\}$ with the switching rule described by a continuous-time Markov chain $\sigma : [0, +\infty) \to \{1, \ldots, 6\}$). To satisfy Assumption 3.1, we chose the Markov chain to the one whose infinitesimal generator is given by

$$\mathcal{Q} = \begin{pmatrix} -0.9500 & 0.4000 & 0.0500 & 0.1000 & 0.1000 & 0.3000 \\ 0.1500 & -0.8500 & 0.4000 & 0.1000 & 0.1000 & 0.1000 \\ 0.2000 & 0.1000 & -0.9000 & 0.1000 & 0.2000 & 0.3000 \\ 0.1000 & 0.2000 & 0.2000 & -0.7000 & 0.1000 & 0.1000 \\ 0.2000 & 0.3000 & 0.1000 & 0.0500 & -0.9500 & 0.3000 \\ 0.1600 & 0.1400 & 0.2000 & 0.3000 & 0.1000 & -0.9000 \end{pmatrix},$$

and to satisfy Assumption 3.1, we chose a group of graphs $\{\mathcal{G}_1, \ldots, \mathcal{G}_6\}$ which are respectively characterized by a group of edge sets $\{\mathcal{E}_1, \ldots, \mathcal{E}_6\}$ with $\mathcal{E}_1 = \{(1,3)\}$, $\mathcal{E}_2 = \{(2,5)\}$, $\mathcal{E}_3 = \{(1,2),(1,3)\}$, $\mathcal{E}_4 = \{(1,2),(1,4),(2,5)\}$, $\mathcal{E}_5 = \{(1,3),(2,5),(3,4)\}$, $\mathcal{E}_6 = \{(1,3),(4,5)\}$. The Markov chain given above is obviously egordic and an invariant measure can be calculated as $\pi = (0.1443, 0.2000, 0.1882, 0.1652, 0.1132, 0.1891)$; a sample path of the Markov chain is shown in Figure 2. For simulation, we chose $s(t) = \frac{1}{t}$ and the initial state as $(1, -4, -2, 3, 4)^T$. The time evolution of the tracking errors are given in Figure 3, which shows the validating of our proposed method.

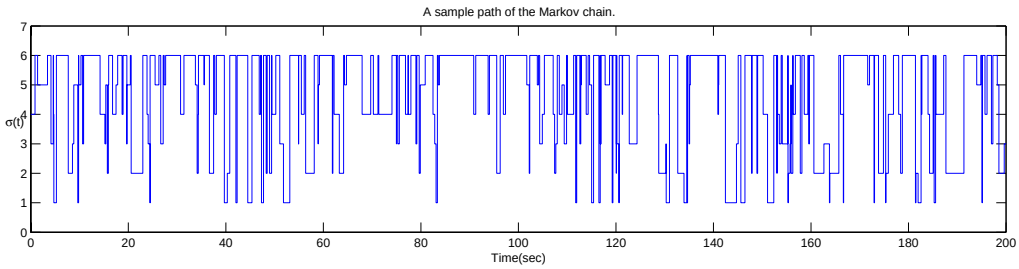

Fig. 2. A sample path of the Markov chain for simulation.

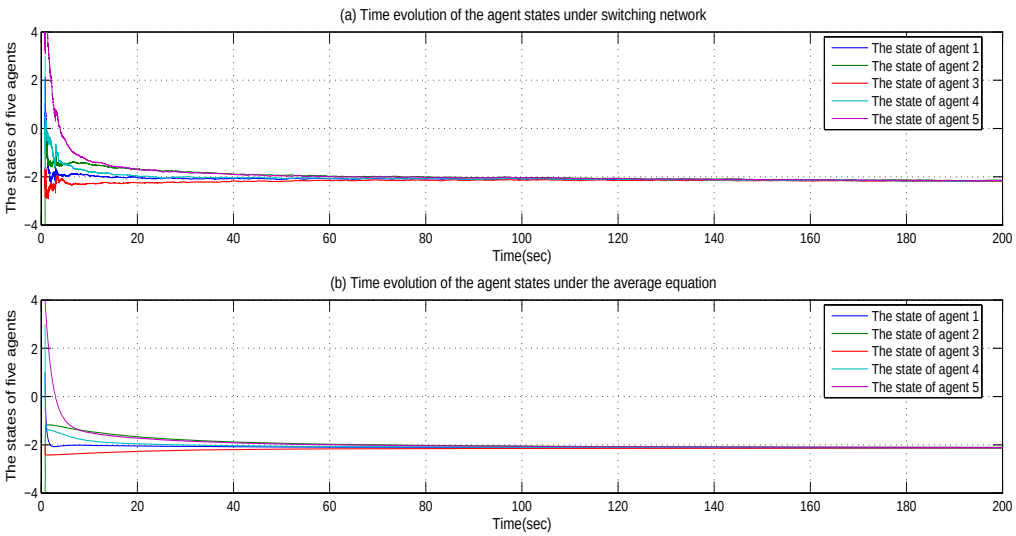

Fig. 3. Time evolution of the agent states under (a) switching network and (b) average equation.

ACKNOWLEDGEMENT

This work is supported by the NNSF of China (61304161, 61663026, 61473098, 61603175, 61563033, 11361043).

(Received August 29, 2016)

REFERENCES

[1] D. P. Bertsekas and J. N. Tsitsiklis: Neuro-Dynamic Programming. Athena Scientific, Belmont 1996.

[2] S. Boyd and L. Vandenberghe: Convex Optimization. Cambridge University Press, 2004. DOI:10.1017/cbo9780511804441

[3] L. C. Evans: Partial Differential Equations. Second edition. American Math Society, 2010.

[4] D. Feijer and F. Paganini: Stability of primal-dual gradient dynamics and applications to network optimization. Automatica *46* (2010), 1974–1981. DOI:10.1016/j.automatica.2010.08.011

[5] D. Freedman: Markov Chains. Springer-Verlag, New York 1983. DOI:10.1007/978-1-4612-5500-0

[6] B. Gharesifard and J. Cortes: Distributed continuous-time convex optimization on weight-balanced digraphs. IEEE Trans. Automat. Control *59* (2014), 781–786. DOI:10.1109/tac.2013.2278132

[7] F. G. Golshtein and N. V. Yakov: Modified Lagrangians and Moonotone Maps in Optimization. Wiley, New York 1996.

[8] K. J. Kushner: Approximation and Weak Convergence Methods for Random Processes with Applications to Stochastic Systems Theory. The MIT Press, London 1984.

[9] S. Liu, Z. Qiu, and L. Xie: Continuous-time distributed convex optimization with set constraints. In: Proc. 19th IFAC World Congress, Cape Town 2014, pp. 9762–9767. DOI:10.3182/20140824-6-za-1003.01377

[10] Q. Liu and J. Wang: A second-order multi-agent network for bounded-constrained distributed optimization. IEEE Trans. Automat. Control *60* (2015), 3310–3315.

[11] Y. Lou, Y. Hong, and S. Wang: Distributed continuous-time approximate projection protocols for shortest distance optimization problems. Automatica *69* (2016), 289–297. DOI:10.1016/j.automatica.2016.02.019

[12] J. Lu and C. Y. Tang: Zero-gradient-sum algorithms for distributed convex: the continuous-time case. IEEE Trans. Automat. Control *57* (2012), 2348–2354. DOI:10.1109/tac.2012.2184199

[13] D. Mateos-Nunez and J. Cortes: Noise-to-state exponential stable distributed convex optimization on weight-balanced digraphs. SIAM J.n Control Optim. *54* (2016), 266–290. DOI:10.1137/140978259

[14] 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

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

[16] G. V. Pavliotis and A. M. Stuart: Multiscale methods: averaging and homogenization. Springer-Verlag, New York 2008.

[17] W. Ni, Xiaoli Wang, and Chun Xiong: Leader-following consensus of multiple linear systems under switching topologies: an averaging method. Kybernetika *48* (2012), 1194–1210.

[18] W. Ni, X. Wang, and C. Xiong: Consensus controllability, observability and robust design for leader-following linear multi-agent systems. Automatica *49* (2013), 2199–2205. DOI:10.1016/j.automatica.2013.03.028

[19] W. Ni, D. Zhao, Y. Ni, and X. Wang: Stochastic averaging approach to leader-following consensus of linear multi-agent systems. J. Franklin Inst. *353* (2016), 2650–2669. DOI:10.1016/j.jfranklin.2016.05.020

[20] R. D. Nowak: Distributed EM algorithms for density estimation and clustering in sensor networks. IEEE Trans. Signal Process. *51* (2003), 2245–2253. DOI:10.1109/tsp.2003.814623

[21] K. Tanabe: A geometric method in nonlinear programming. J. Optim. Theory Appll. *30* (1980), 181–210. DOI:10.1007/bf00934495

[22] B. Touri and B. Gharesifard: Continuous-time distributed convex optimization on time-varying directed networks. In: 54th IEEE Conference on Decision and Control, Osaka 2015, pp. 724–729. DOI:10.1109/cdc.2015.7402315

[23] J. Wang and N. Elia: Control approach to distributed optimization. In: 48th Annual Allerton Conference on Communication, Control, and Computing 2010, pp. 557–561. DOI:10.1109/allerton.2010.5706956

[24] J. Wang and N. Elia: A control perspective for centralized and distributed convex optimization. In: 50th IEEE Conference on Decision and Control and European Control Conference, Orlando 2011, pp. 3800–3805.

[25] X. Wang, P. Yi, and Y. Hong: Dynamic optimization for multi-agent systems with external disturbances. Control Theory Technol. *12* (2014), 132–138. DOI:10.1007/s11768-014-0036-y

[26] P. Yi, and Y. Zhang, and Y. Hong: Potential game design for a class of distributed optimisation problems. J. Control Decision *1* (2014), 166–179.

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

[28] P. Yi, Y. Hong, and F. Liu: Initialization-free distributed algorithms for optimal resource allocation with feasibility constraints and its application to economic dispatch of power systems. Automatica *74* (2016), 259–269.

[29] X. Zeng, P. Yi, and Y. Hong: Distributed continuous-time algorithm for constrained convex optimizations via nonsmooth analysis approach. arXiv:1510.07386v2, 2016.

Wei Ni, Corresponding author. Numerical Simulation and High-Performance Computing Laboratory, School of Science, Nanchang University, Nanchang 330031. P. R. China.

e-mail: niw@amss.ac.cn

Xiaoli Wang, School of Information Science and Engineering, Harbin Institute of Technology at Weihai, Weihai 264209. P. R. China.

e-mail: xiaoliwang@amss.ac.cn