A Multi-Scale Method for Distributed Convex Optimization with Constraints

Wei Ni\textsuperscript{1} · Xiaoli Wang\textsuperscript{2}

Received: 2 January 2021 / Accepted: 19 November 2021 / Published online: 6 January 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract
This paper proposes a multi-scale method to design a continuous-time distributed algorithm for constrained convex optimization problems by using multi-agents with Markov switched network dynamics and noisy inter-agent communications. Unlike most previous work which mainly puts emphasis on dealing with fixed network topology, this paper tackles the challenging problem of investigating the joint effects of stochastic networks and the inter-agent communication noises on the distributed optimization dynamics, which has not been systemically studied in the past literature. Also, in sharp contrast to previous work in constrained optimization, we depart from the use of projected gradient flow which is non-smooth and hard to analyze; instead, we design a smooth optimization dynamics which leads to easier convergence analysis and more efficient numerical simulations. Moreover, the multi-scale method presented in this paper generalizes previously known distributed convex optimization algorithms from the fixed network topology to the switching case and the stochastic averaging obtained in this paper is a generalization of the existing deterministic averaging.

Keywords Distributed convex optimization · Multi-scale method · Multi-agent systems · Stochastic averaging · Backward Kolmogorov equation

Communicated by Xiaoqi Yang.

Wei Ni
niw@amss.ac.cn

Xiaoli Wang
xiaoliwang@amss.ac.cn

\textsuperscript{1} School of Sciences, Nanchang University, Nanchang 330031, China

\textsuperscript{2} School of Information Science and Engineering, Harbin Institute of Technology at Weihai, Weihai 264209, China
1 Introduction

The research of convex optimization by using multi-agent systems is a hot topic in recent decades. This problem typically takes the form of minimizing the sum of functions and is usually divided into subtasks of local optimizations, where each local optimization subtask is executed by one agent and the cooperation among these agents makes these local algorithms compute the optimal solution in a consensus way. Informally, convex optimization algorithms constructed in this way are usually termed as distributed convex optimization (DCO), which models a broad array of engineering and economic scenarios and finds numerous applications in diverse areas such as operations research, network flow optimization, control systems and signal processing; see [1,7,9,21–23,33] and references therein for more details.

Usually, DCO takes advantages of the consensus algorithms in multi-agent systems and gradient algorithms in convex optimization. The idea of combining them was proposed early in 1980s by Tsitsiklis et al. in [35] and re-examined recently in the context of DCO in [1,7,21,23]. Most DCO algorithms in the earlier development were discrete-time, with the distributed gradient descent strategy [21,23,33] being the most popular. Various extensions of the distributed gradient descent were then proposed, such as push-sum-based approach [22], incremental gradient procedure [34], conditional gradient scheme [6], proximal method [29], fast distributed gradient strategy [15,31], and non-smooth analysis-based technique [45], just to name a few. Additionally, using local gradients is rather slow and the community has moved toward using gradient estimation [31,41] and stochastic gradient [33,44]. Further research directions were then followed by taking optimization constraints into consideration. Generally, DCO with constraints has proceeded along two research lines, namely projected gradient strategy and primal-dual scheme. The projected gradient strategy designed the optimization algorithms by projecting the gradient into the constraint set and extended it by including a consensus term [7,19,33,37]. The primal-dual scheme introduced equality and inequality multipliers and designed for the extra dual dynamics [5,43,46] in which a projection onto positive quadrant is usually included [5,9,37,43] so that the inequality-multiplier stays positive. Therefore, both research lines dealing with optimization constraints above are projection-dependent.

While projection-dependent algorithms were widely used in DCO, they require the optimization constraints to have a relatively simple form so that the projections can be computed analytically. To overcome this difficulty, this paper pursues a new method to design a novel DCO algorithm by avoiding projection. Our method is built on the primal-dual setup by introducing Lagrange multipliers. We modify the classical projection-based dynamics for the inequality-multiplier by utilizing the technique of mirror descent [24,32] and design a projection-free multiplier-dynamics which is smooth. In conclusion, compared with most existing constrained DCO algorithms, our method avoids projection and thus reduces the difficulties of convergence analysis and iterative computation.

Aside from the difficulty associated with optimization constraints, the second challenge in DCO problem ties with stochastic networks and inter-communication noises. This challenge, together with optimization constraints, jointly makes the DCO problem difficult to analyze, and therefore relatively few results were reported. Nedic [22]
considered DCO over deterministic and uniformly strongly connected time-varying networks without considering communication noises and optimization constraints; furthermore, their algorithm needed a strong requirement that each node knows its out-degree at all times. The DCO problem with optimization constraints over time-varying graphs was investigated in [40] by using the epigraph form, but communication noises were not considered there. The work in [18] also investigated the consensus-based DCO algorithm by using a random graph model where the communication link availability is described by a stochastic process, but leaving challenging issues of optimization constraints and communication noise untouched.

To tackle the above-mentioned difficulties and to contribute to the existing literature, this paper proposes a multi-scale method for constrained DCO problem over Markov switching networks under noisy communications. Unlike most existing DCO algorithms which are discrete-time, we study this problem in a continuous-time framework because the classical tools of Ito formula, backward Kolmogorov equation and ergodic theory in stochastic analysis can be used and the elegant Lyapunov argument in optimization theory [9] can be invoked. Recently, we established in [25] a new technique of stochastic averaging (SA) for unconstraint DCO, where the idea of averaging was previously explored by us to handle the switching networks of the multi-agent systems, with the deterministic version being presented in [26, 27] and the stochastic version in [28]. Compared with [25], the present work considers a more general case of constrained optimization which is more challenging.

Although the SA viewpoint for multi-agent systems has been indicated in [28], its theoretical clarification and design details were not provided there. In this paper, we generalize the SA principle in [28] to the DCO problem and propose a multi-scale-based design procedure for the SA. We begin with the intuition behind our approach. Our multi-scale method borrows the idea of the slave principle in Synergetics [13] which was initially proposed by German physician Haken in 1970s. According to this principle, the system variables are classified into fast and slow ones, where the slow variables dominate the system evolution and characterize the ordering degree of the system. Therefore, it is necessary to eliminate the fast variables and obtain an equation for the slow variables only. This equation is called the principle equation and it can be viewed as an approximate description of the system. The method of eliminating the fast variables in physics is termed as Born–Oppenheimer approximation. In this paper, the continuous-time Markov chain characterizing the time-varying networks is regarded as the fast variable and, in contrast, the states of the optimization multi-agent system are considered as slow variables. To distinguish the fast and slow variables, two time-scales for them are introduced. In this paper, we propose a concrete scheme to eliminate the fast variable by resorting to the tool of multi-scale analysis introduced by Pavliotis [30] and obtain an averaged SDE which acts as an approximation to the original switching stochastic differential equation (SDE). In this sense, the effect of network switching on optimization dynamics is eliminated, and thus the DCO under switching networks is, in fact, reduced to that under fixed case.

We mention two benefits of our multi-scale method. The first benefit in comparison with those dealing with random networks (see, e.g., [18, 21]) lies in its generalizability. The widely used DCO algorithms over random networks in [18, 21] and the online DCO algorithm in [14] built their analysis on the product theory of stochastic matrices
(c.f. [39]): the products of stochastic matrices converge to a rank one matrix (see Lemmas 4-7 in [18]). This theory was used to analyze the convergence of their optimization algorithm which is driven by a chain of stochastic matrices and an inhomogeneous gradient term (see Section V in [18]). However, due to technicalities involved, this method is hard to generalize to include optimization constraints or communication noises since otherwise the resulting matrices are not stochastic matrices. Our method does not have this limitation, instead it can treat the optimization constraints, communication noises and stochastic networks into a unified framework. As the second benefit, our method reduces the optimization algorithm in stochastic networks to that in fixed network (see Theorem 6.1) and establishes an approximation relationship between the two algorithms (see Theorem 6.2). Therefore, the SA method in this paper can help generalize existing DCO algorithms from fixed network to stochastic networks.

To sum up, the contributions of our paper are as follows.

- Firstly, we propose a novel method of SA for the DCO problem and addresses in a unified framework the challenging issues of the optimization constraints, communication noises and stochastic networks.
- Secondly, the SA method can help generalize some DCO algorithms from fixed network to switching case since it converts the latter to the former and establishes an approximation relationship between them.
- Thirdly, the DCO algorithm in this paper is projection-free and thus has the advantages of removing the difficulty of computing projection and rendering the resulting optimization dynamics to be smooth so that algorithm analysis and simulation become relatively easy.
- Lastly, the multi-scale method used in this paper has the ability to generalize the averaging method from the deterministic case [26, 27] to the stochastic case in the present form, and it can also provide a theoretical justification for our original vision of SA for multi-agent systems [28].

2 Preliminaries

A. Notations. For a vector $a$, its $i$-th component is denoted by $[a]_i$ or $a_i$. By $a \prec 0$ ($a \preceq 0$), we mean that each entry of $a$ is less than (less than or equal to) zero. Letting $a = (a_1, \ldots, a_n)^T \in \mathbb{R}^n$ and $b = (b_1, \ldots, b_n)^T \in \mathbb{R}^n$, we define $a \odot b = (a_1b_1, \ldots, a_nb_n)^T$ and $a \oslash b = (a_1/b_1, \ldots, a_nb_n)^T$. The notation $1_n$ denotes an $n$-dimensional vector with each entry being 1. We use $\mathbb{R}_+^n$ ($\mathbb{R}^n_{++}$) to denote the set of $n$-dimensional vectors with nonnegative (positive) components. For vectors $\alpha_1, \ldots, \alpha_m$, the notation $\text{col}\{\alpha_i\}_{i=1}^m$ denotes a new vector $(\alpha_1^T, \ldots, \alpha_m^T)^T$. For matrices $M_1, \ldots, M_m$, we use $\text{diag}\{M_i\}_{i=1}^m$ to denote block diagonal matrix with $i$-th block being $M_i$. The inner product between matrices is denoted as $A : B = \text{tr}(A^T B) = \sum_{i,j} a_{ij}b_{ij}$, where $\text{tr}$ denotes the matrix trace. For a map $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ which is differentiable at $x$, we use $\nabla g(x)$ to denote the matrix whose rows are the gradients of the corresponding entries in the vector $g(x)$. Given a continuously-differentiable, strictly convex function $\psi$ defined on a closed convex set $\Omega$, the
Bregman divergence associated with $\psi$ for points $a, b \in \Omega$ is defined as $D_\psi(a, b) = \psi(a) - \psi(b) - (a - b)^T \nabla \psi(b)$.

**B. Graph Theory.** Consider 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. The graph considered in this paper is undirected in the sense that the edges $(i, j)$ and $(j, i)$ in $\mathcal{E}$ are considered to be the same. The set of neighbors of node $i$ is denoted by $\mathcal{N}_i = \{j \in \mathcal{V} : (i, j) \in \mathcal{E}, j \neq i\}$. We use the symbol $\cup$ to denote the graph union.

We say that a collection of graphs is jointly connected if the union of its members $\mathcal{G}_1, \cdots, \mathcal{G}_S$ is jointly connected if and only if the matrix $L_1 + \cdots + L_S$ has a simple zero eigenvalue, where $L_1, \cdots, L_S$ are, respectively, the Laplacians of the graphs $\mathcal{G}_1, \cdots, \mathcal{G}_S$; see [26] for more details.

### 3 Problem Formulation

Consider an optimization problem on a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Each agent $i \in \mathcal{V}$ has a local cost function $f_i : \mathbb{R}^n \to \mathbb{R}$ and a group of local inequality constraints $g_{ij}(x) \leq 0$, $j = 1, \ldots, r_i$ and equality constraints $h_{ik}(x) = 0$, $k = 1, \ldots, s_i$, where $r_i$ and $s_i$ are nonnegative integers. If there is no constraints for agent $i$, one simply sets corresponding constraint functions to be zero. The total cost function of the network is given by sum of all local functions, and the optimization is to minimize the global cost function of the network while satisfying $N$ group of local constraints, given explicitly as follows,

$$\begin{array}{ll}
\mathcal{P} : & \text{minimize} \quad \tilde{f}(x) = \sum_{i=1}^{N} f_i(x), \\
& \text{subject to} \quad g_i(x) \leq 0, \\
& \quad h_i(x) = 0, i = 1, \ldots, N, \\
\end{array} \tag{1}$$

where $f_i : \mathbb{R}^n \to \mathbb{R}$, $g_i = (g_{i1}, \cdots, g_{ir_i})^T : \mathbb{R}^n \to \mathbb{R}^{r_i}$, and $h_i = (h_{i1}, \cdots, h_{is_i})^T : \mathbb{R}^n \to \mathbb{R}^{s_i}$, are, respectively, the local cost, inequality constraint and equality constraint on node $i$. In what follows, we assume that $f_i$, $g_i$, $i = 1, \cdots, N$ are convex and twice differentiable, and $h_i$, $i = 1, \cdots, N$ are affine, so that the optimization problem (1) is convex.

Let $x^*$ be an optimal solution, if exists, to the problem (1). If additional assumptions on the constraint functions, called constrained qualifications, are satisfied, then the following classical Karush–Kuhn–Tucker (KKT) conditions hold at the minimizer $x^*$: there exist multipliers $\lambda^*_{ij} \in \mathbb{R}$ and $\nu^*_{ij} \in \mathbb{R}$ such that, for $i = 1, \cdots, N$,

$$\begin{align}
&g_{ij}(x^*) \leq 0, j = 1, \cdots, r_i, \tag{2a} \\
h_{ij}(x^*) = 0, j = 1, \cdots, s_i, \tag{2b} \\
&\lambda^*_{ij} \geq 0, j = 1, \cdots, r_i, \tag{2c} \\
&\lambda^*_{ij} g_{ij}(x^*) = 0, j = 1, \cdots, r_i, \tag{2d} \\
&\sum_{i=1}^{N} \nabla f_i(x^*) + \sum_{i=1}^{N} \sum_{j=1}^{r_i} \lambda^*_{ij} \nabla g_{ij}(x^*) + \sum_{i=1}^{N} \sum_{j=1}^{s_i} \nu^*_{ij} \nabla h_{ij}(x^*) = 0. \tag{2e}
\end{align}$$
A widely used constrained qualification is the Slater’s constrained qualification (SCQ): there exists $x \in \mathbb{R}^n$ such that $g_i(x) < 0$ and $h_i(x) = 0$ for $i = 1, \ldots, N$. In other words, assuming SCQ, “$x^*$ solves (P)” $\Rightarrow$ “$\exists$ a set of $(\lambda_{ij}^*, v_{ij}^*)$ together with $x^*$ solving (KKT)”. Furthermore, for convex problem, this implication is bidirectional. Refer to [2] for details.

While SCQ ensures the existence of multipliers satisfying (KKT), it does not guarantee uniqueness. Closely tied to this direction is the linear independence constraint qualification (LICQ), which is stronger than SCQ. Using $\nabla g_i(x^*)$ to denote the sub-matrix of $\nabla g_i(x^*)$ given by rows with indices in $J_i(x^*) = \{j | g_{ij}(x^*) = 0\}$, the LICQ is defined as

$$\text{rank} \left[ \begin{array}{c} \nabla h_i(x^*) \\ \nabla g_i(x^*) \end{array} \right] = s_i + |J_i(x^*)|, \quad i = 1, \ldots, N,$$

(3)

here $| \cdot |$ denotes the set cardinality. By assuming LICQ, one obtains a result [36] on the existence and uniqueness of multipliers satisfying (2),

“$x^*$ solves (P)” $\Rightarrow$ “$\exists$ a unique $(\lambda_{ij}^*, v_{ij}^*)$ together with $x^*$ solving (KKT)”. (4)

This direction is bidirectional if the problem is convex. To ensure uniqueness of multipliers, we make the following assumption:

**Assumption 1** The LICQ defined in (3) is satisfied.

Also, we assume the existence of optimal solutions without giving explicit conditions due to space limitation; interested readers can refer to [2]. We further assume strict convexity on at least one function among $\{f_1, \ldots, f_i\}$, so that there exists at most one global optimal solution to the problem (1). With these assumptions, the unique optimal solution is denoted by $x^* \in \mathbb{R}^n$.

### 4 Distributed Optimization Dynamics

We use $N$ agents $\{1, \ldots, N\}$ to solve the convex optimization problem (1) in a distributed way. Each agent, say $i$, is a dynamical system with states $(x_i(t), \theta_i(t), \lambda_i(t), v_i(t))$, where $x_i \in \mathbb{R}^n$ is the estimation of the optimal solution $x^*$, $\lambda_i \triangleq \text{col}\{\lambda_{i1}, \ldots, \lambda_{ir_i}\} \in \mathbb{R}^{r_i}$ and $v_i \triangleq \text{col}\{v_{i1}, \ldots, v_{ir_i}\} \in \mathbb{R}^{r_i}$ are, respectively, the estimations of the Lagrange multipliers $\lambda_{ij}^* \triangleq \text{col}\{\lambda_{ij_{1}}, \ldots, \lambda_{ij_{r}}\} \in \mathbb{R}^{r_i}$ and $v_{ij}^* \triangleq \text{col}\{v_{ij_{1}}, \ldots, v_{ij_{r}}\} \in \mathbb{R}^{r_i}$, and $\theta_i \in \mathbb{R}^n$ is an auxiliary vector to be specified later. Agent $i$ only exchanges its estimate $x_i$, rather than $\theta_i, \lambda_i, v_i$, with its neighbors. Therefore, the number of variables transmitted across the network is reduced and the communication cost is decreased. As a result, consensus will only be reached for $x_i$.

We consider the general case that the communication is corrupted by noises which lie in a filtered probability space $(\Omega, \mathcal{F}, \mathbb{P}, \{\mathcal{F}_i\}_{t \geq 0})$, where $\{\mathcal{F}_i\}_{t \geq 0}$ is a sequence of increasing $\sigma$-algebras with $\mathcal{F}_\infty \subset \mathcal{F}$ and $\mathcal{F}_0$ containing all the $\mathbb{P}$-null sets in $\mathcal{F}$. In more details, for any two neighbor agents $i, j$ with communication channel $(j, i)$
connecting them, the ideal relative information \((x_j - x_i)\) transmitted in this channel is corrupted by state-dependent noise \(\sigma_{ji} \xi_{ji}(x_j - x_i)\) with \(\xi_{ji}(t) \in \mathbb{R}\) being independent standard white noises adapted to the filtration \(\{\mathcal{F}_t | t \geq 0\}\) and \(\sigma_{ji} \geq 0\) being noise intensity satisfying the following assumption:

**Assumption 2** The noise intensity \(\sigma_{ij}\) has an upper bound \(\sigma_{ij} \leq \kappa\) for some positive constant \(\kappa\), where \(i, j = 1, \cdots, N\).

The above kind of noise indicates that the closer the agents are to each other, the smaller the noise intensities. The noise of this type is multiplicative in nature. While additive noise provides a natural intuition, multiplicative noise model also has its practical background. For example, it can model the impact of quantization error, as well as the effect of a fast-fading communication channel. Also, lossy communication-induced noises and imperfect sample-induced noises are all multiplicative noises. We also note that the issue of adopting multiplicative noise in inter-agent communication has been well explained in existing researches (see for example references [27], [17, Remarks 1-2], [4, 38] and references therein). By adopting this noise model, we design for each agent \(i \in \{1, \cdots, N\}\) the following optimization dynamics

\[
\dot{x}_i = c \sum_{j \in \mathcal{N}_i(t)} (1 + \sigma_{ji} \xi_{ji}) (x_j - x_i) - \nabla f_i(x_i) - \theta_i - \sum_{j \in \mathcal{N}_i(t)} r_{ij} \lambda_{ij} \nabla g_{ij}(x_i), \tag{5a}
\]

\[
\dot{\theta}_i = -c \sum_{j \in \mathcal{N}_i(t)} (1 + \sigma_{ji} \xi_{ji}) (x_j - x_i), \tag{5b}
\]

\[
\dot{\lambda}_{ij} = \frac{\lambda_{ij}}{1 + \eta_{ij} \lambda_{ij}} g_{ij}(x_i), \quad j = 1, \cdots, r_i, \tag{5c}
\]

\[
\dot{v}_{ij} = h_{ij}(x_i), \quad j = 1, \cdots, s_i, \tag{5d}
\]

where \(x_i, \theta_i \in \mathbb{R}^n, \lambda_{ij}, v_{ij} \in \mathbb{R}, \eta_{ij}\) are positive parameters and \(c > 0\) is the coupling strength. As can be seen, with increased number of the agents, the amount of calculation for individual agent remains unchanged if the newly added members are not neighbors of this agent; however, the amount of inter-agent communication increases since there are more agents involved in the communication network.

Equations (5a)–(5b) are motivated by [16, Eq. (3)], which however does not consider optimization constraints, communication noises, and more importantly time-varying network. This series of hard problems are tackled in our paper. Equation (5c) is motivated by [3, Eq. (4)], but differs from [3] since it is distributed by using \(N\) agents on a network and considers the communication noises among agents and the stochastic networks. The dynamics (5d) can be obtained by maximizing the Lagrangian \(\Phi\) in (8) with respect to \(v_{ij}\) via gradient ascent \(\dot{v}_{ij} = \nabla_{v_{ij}} \Phi\).

The algorithm in (5) is designed for time-varying networks. Assume that there are \(S \in \mathbb{N}\) possible graphs \(\{G_1, \cdots, G_S\}\), among which the network structure is switched. We use a continuous-time Markov chain \(\sigma : [0, \infty) \to S := \{1, \cdots, S\}\) to describe this switching. To analyze the stability of Eq. (5), we rewrite it into a switching SDE
as (refer to Appendix A for detailed derivation and implicit definition of symbols in the equation below),

\[
\begin{align*}
    dx &= [-cL_\sigma(t)x - \theta - \nabla F(x) - \lambda \odot \nabla G(x) - v \odot \nabla H(x)]dt + cM_\sigma(t)dw, \\
    d\theta &= cL_\sigma(t)x dt - cM_\sigma(t)dw, \\
    d\lambda &= [\lambda \odot (1 + \eta \odot \lambda)] \odot G(x)dt, \\
    dv &= H(x)dt.
\end{align*}
\]

Note that we have assumed the existence and uniqueness of an optimal solution \(x^*\). Defining \(x^* = 1_N \otimes x^*\), it follows from (4) and Assumption 1 that there is a unique pair of \((\lambda^*, v^*)\) satisfying (2), where \(\lambda^*\) and \(v^*\) are, respectively, stacked vectors of \(\lambda^*_{ij}\) and \(v^*_{ij}\) in (4). These give rise to another vector \(\theta^*\) defined by

\[
\theta^* + \nabla F(x^*) + \lambda^* \odot \nabla G(x^*) + v^* \odot \nabla H(x^*) = 0. \tag{7}
\]

Obviously, \((x^*, \theta^*, \lambda^*, v^*)\) satisfies (6) and consequently is an equilibrium of (6) (note that stochastic disturbances vanish at this equilibrium).

In the rest of this paper, we will utilize the averaging method used in [25,28] to analyze the stability of the equilibrium \((x^*, \theta^*, \lambda^*, v^*)\) for (6).

5 Distributed Optimization Under Fixed Network

In this section, we assume that the network is fixed, so that the time-varying graph Laplacian \(L_\sigma(t)\) and the diffusion term \(M_\sigma(t)\) in the dynamics (6) are replaced with fixed ones \(L\) and \(M\), respectively. The following assumption is made to ensure the positiveness of a parameter which is specified later and used in the proof of the stability for our optimization dynamics.

Assumption 3 The coupling strength \(c\) for fixed network satisfies \(0 < c < \frac{2}{3}\kappa^{-2}\).

We are now in a position to present the main result of this section. The following theorem states that the trajectory of our optimization dynamics converges to its equilibrium \((x^*, \theta^*, \lambda^*, v^*)\) under appropriate conditions. While this result is obtained for fixed network, the multi-scale method developed in next section can transform the analysis of the optimization algorithm under switching networks to that under fixed one.

Theorem 5.1 For the constrained optimization problem (1), suppose there are \(N\) agents with each agent dynamics given by (5), and they are connected across a fixed network (the graph Laplacian and diffusion term in (6) are now denoted as \(L\) and \(M\), respectively). Under Assumptions 1–3 with \(\kappa \leq \sqrt{\frac{\lambda_2 L}{2}}\) with \(\lambda_2\) the smallest nonzero eigenvalue of the graph Laplacian \(L\), then for any trajectory of (6) with initial conditions \(\lambda_{ij}(0) > 0\) and \(\sum_{i=1}^N \theta_i(0) = 0\), one has \(\lim_{t \to \infty} \|x_i(t) - x^*\| = 0\) for \(i = 1, \ldots, N\) almost surely.
Proof We use the method of Lyapunov function to prove the stability, with the key to construct an appropriate Lyapunov function. Here we only give a proof skeleton, with more details being put in Appendices B, C, and D.

1° Construct a Lyapunov candidate \( V(x, \theta, \lambda, v) = V_1 + V_2 + V_3 + V_4 \) with

\[
V_1 = \frac{1}{2} \| x - x^* \|^2 + \frac{1}{2} \| (x - x^*) + (\theta - \theta^*) \|^2,
\]

\[
V_2 = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{r_i} \eta_{ij}(\lambda_{ij} - \lambda_{ij}^*)^2,
\]

\[
V_3 = \sum_{(i,j) \in \Omega} D_\psi(\lambda_{ij}, \lambda_{ij}^*) + \sum_{(i,j) \notin \Omega} (\lambda_{ij} - \lambda_{ij}^*)^2,
\]

\[
V_4 = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{s_i} (v_{ij} - v_{ij}^*)^2,
\]

where \( \Omega = \{(i,j) | \lambda_{ij}^* \neq 0 \} \) and \( D_\psi(\lambda_{ij}, \lambda_{ij}^*) \geq 0 \) is the Bregman divergence between \( \lambda_{ij} \) and \( \lambda_{ij}^* \) with respect to \( \psi(x) = x \ln x \). Therefore, \( V \geq 0 \), and also \( V(x, \theta, \lambda, v) = 0 \) if and only if \( (x, \theta, \lambda, v) = (x^*, \theta^*, \lambda^*, v^*) \).

2° We now calculate the action of the operator \( \mathcal{A} \) on the function \( V \) (Recall that, for an SDE \( dx = adt + bdw, AV = \nabla V \cdot a + \frac{1}{2} \text{tr}(b^T \nabla^2 V b) \)). Defining a Lagrangian \( \Phi : \mathbb{R}^{nN} \times \mathbb{R}^{nN} \times \mathbb{R}_+^N \times \mathbb{R}^T \to \mathbb{R} \) as

\[
\Phi(x, \theta, \lambda, v) = \Psi + \sum_{i=1}^{N} \sum_{j=1}^{r_i} \lambda_{ij} g_{ij}(x_i) + \sum_{i=1}^{N} \sum_{j=1}^{s_i} v_{ij} h_{ij}(x_i),
\]

\[
\Psi(x, \theta) = \sum_{i=1}^{N} f_i(x_i) + (x - x^*)^T \theta + \lambda^T \mathcal{L} x, \quad \lambda = \frac{1}{2}(c - \frac{3}{2}c^2 \kappa^2),
\]

we can show in Appendix B that

\[
\mathcal{A} V \leq \Phi(x^*, \theta, \lambda, v) - \Phi(x, \theta^*, \lambda^*, v^*) - (h \lambda_2^G - 1)(x - x^*)^T (x - x^*),
\]

where \( \lambda_2^G \) is the smallest nonzero eigenvalue of the connected graph \( G \). Noting that \( (x^*, \lambda^*, v^*) \) satisfies (2), we show in Appendix C that the following saddle point condition holds

\[
\Phi(x^*, \theta, \lambda, v) \leq \Phi(x^*, \theta^*, \lambda^*, v^*) \leq \Phi(x, \theta^*, \lambda^*, v^*).
\]

Therefore, \( \mathcal{A} V \leq 0 \) if \( h \lambda_2^G - 1 > 0 \), which is guaranteed by \( \kappa \leq \sqrt{\lambda_2^G / 2} \).

3° Letting \( \mathcal{A} V = 0 \), we show in Appendix D that \( (x, \theta, \lambda, v) = (x^*, \theta^*, \lambda^*, v^*) \). The application of the stochastic version of the Lasalle’s invariance principle in \([20]\) yields that the equilibrium \( (x^*, \theta^*, \lambda^*, v^*) \) of system (5) is asymptotically stable almost surely.

\[\square\]
6 Multi-Scale Analysis of Distributed Optimization Dynamics

This section analyzes the optimization dynamics under switching networks by proposing a multi-scale analysis. This method can reduce the analysis of optimization algorithm under switching topologies to that under a fixed one.

The switching feature encoded in $\sigma(t)$ makes difficult the convergence analysis of system (6). To cope with this difficulty, we adopt an idea of SA proposed in our earlier works [25, 28]. The basic idea is to approximate in an appropriate sense the switching system (6) using a non-switching system, called the average system. A detailed construction of such an average system is provided in this section by resorting to the multi-scale analysis. Noting that adding or deleting even one edge in the graph may result in the change of the network structure, and also noting that the graph of the network is large scale (i.e., the numbers of nodes and edges are extremely large), the network structure changes more frequently than the evolution of the optimization dynamics on the nodes. That is, the state $\sigma(t)$ of the network structure changes faster than the state of the optimization $\sigma(\alpha)$. We use a small parameter $\alpha > 0$ to rescale $\sigma(t)$ so that $\sigma(t/\alpha)$ is a fast process and $\sigma(\alpha)$ is a slow process. Correspondingly, the SDE (6) under the above re-scaling admits the following form

$$
\begin{align}
\frac{d\mathbf{x}^\alpha}{dt} &= [-c\mathcal{L}_{\sigma(t)}\mathbf{x}^\alpha - \mathbf{\theta}^\alpha - \nabla F(\mathbf{x}^\alpha) - \lambda^\alpha \odot \nabla G(\mathbf{x}^\alpha) - \mathbf{v}^\alpha \odot \nabla H(\mathbf{x}^\alpha)]dt + c\mathcal{M}_{\sigma(t)}d\mathbf{w}, \\
\frac{d\mathbf{\theta}^\alpha}{dt} &= c\mathcal{L}_{\sigma(t)}\mathbf{x}^\alpha, \\
\frac{d\lambda^\alpha}{dt} &= \left[\lambda^\alpha \odot (1 + \eta \odot \lambda^\alpha)\right] \odot G(\mathbf{x}^\alpha), \\
\frac{dv^\alpha}{dt} &= H(\mathbf{x}^\alpha).
\end{align}
$$

The time re-scaling is crucial in the method of SA and its role will be seen later. Some assumptions on the Markov chain $\sigma(t)$ are now given below.

6.1 Assumptions on The Markov Switching Network

It is known that the statistics of a Markov chain defined over $\mathbb{S} = \{1, \ldots, S\}$ is identified by an initial probability distribution $\pi_0 = [\pi_{01}, \ldots, \pi_{0S}]^T$ with $\pi_{0i} = \mathbb{P}[\sigma(0) = i]$ and by a Metzler matrix $Q = (q_{ij})_{S \times S} \in \mathbb{R}^{S \times S}$. This matrix is also called the infinitesimal generator of the Markov chain and it describes the transition probability as

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

where $\Delta t > 0$, $q_{ij} \geq 0$ ($i \neq j$) is the transition rate from state $i \in \mathbb{S}$ at time $t$ to state $j \in \mathbb{S}$ at time $t + \Delta t$, and $q_{ii} = -\sum_{j \neq i} q_{ij}$, $\lim_{\Delta t \to \infty} o(\Delta t)/\Delta t = 0$ (Note that $|q_{ij}| < \infty$ since $\sigma(t)$ is a finite state Markov chain [11, pp.150–151]). More specifically, at time $t$ the state of the Markov chain is determined according to the probability distribution $\pi(t) = (\pi_1(t), \ldots, \pi_S(t))$ with $\pi_s(t)$ being the probability that at time $t$ the Markov system is in the state $s \in \mathbb{S}$. The normalization condition

$$
\sum_s \pi_s(t) = 1.
$$
$\sum_{s=1}^{S} \pi_s(t) = 1$ is usually assumed. Letting $\Delta t \to 0$, the infinitesimal form of the Markov dynamics reads as $\dot{\pi}(t) = \sum_{i=1}^{S} \pi_i(t) q_{is}, s = 1, \cdots, S$. In a compact form, the distribution $\pi(t)$ for $\sigma(t)$ obeys

$$\dot{\pi}(t) = Q^T \pi(t). \quad (13)$$

Since we are interested in the asymptotic behavior of the system, we will assume that the probability distribution $\pi(t)$ of $\sigma(t)$ is stationary, and it is denoted by $\pi$, which is defined as

$$Q^T \pi = 0, \quad \sum_{i=1}^{S} \pi_i = 1, \quad \pi_i > 0. \quad (14)$$

The existence of the stationary distribution $\pi$ satisfying (14) can be guaranteed by the ergodicity of $\sigma$, namely, all graphs $\{G_1, \cdots, G_S\}$ can be visited infinitely often under the switching $\sigma$. The joint connectivity of the network is also needed for later analysis and it is also assumed here.

**Assumption 4** The finite-state Markov process $\sigma(t)$ describing the switching networks has a stationary probability distribution $\pi = (\pi_1, \cdots, \pi_N)^T$ satisfying (14), and the union graph $\bigcup_{s \in S} G_s$ is connected.

Due to Assumption 4, the eigenvalues $\tilde{\lambda}_1 < \tilde{\lambda}_2 \leq \cdots \leq \tilde{\lambda}_N$ of the matrix $\tilde{L} := \sum_{i=1}^{S} L_i$ has a simple zero eigenvalue $\tilde{\lambda}_1 = 0$. Denote $\pi_{\min} = \min\{\pi_1, \cdots, \pi_S\}$ and $\pi_{\max} = \max\{\pi_1, \cdots, \pi_S\}$. For later use, we modify Assumption 3 in fixed topology to generalize to switching topology as follows.

**Assumption 3’** The coupling strength $c$ under the switching network satisfies $0 < c < \frac{2}{3}(\pi_{\min}/\pi_{\max})\kappa^{-2}$.

The following lemma, which will be used later, is a slight modification of [10, Lemma 4.2], and it can be proved similarly as in [10].

**Lemma 6.1** Suppose that $V(t)$ is $\mathcal{F}$-measurable and $\mathbb{E}\{V(t)1_{\{|\sigma(t)|=i\}}\} \exists$, where $1_{|\sigma(t)|=i}$ is the indicator function of the event $\{|\sigma(t)| = i\}$. Then $\mathbb{E}\{V(t) \cdot d(1_{|\sigma(t)|=i})\} = \frac{1}{\alpha} \sum_{j=1}^{S} q_{js} \mathbb{E}\{V(j)(t)\}dt + o(dt)$ holds for each $s \in S$, where $V(i)(t) = V(t)1_{|\sigma(t)|=i}$.

### 6.2 Stochastic Averaging Method for Switching Networks

The properties of the solutions to the SDE (12) are included in its backward Kolmogorov equation which is a partial differential equation determined by the infinitesimal of (12). Although the analytic solutions to the backward Kolmogorov equation are hard to obtain, we focus on those solutions which are Taylor series in term of $\alpha$. We use the first term in the series as an approximate solution. It will be shown that this approximate solution satisfies another backward Kolmogorov equation, which is called as the averaged backward Kolmogorov equation. This averaged backward Kolmogorov equation is nothing but the one whose operator is the average of infinitesimals of all the subsystems in the switched system (12). Corresponding to
this average backward Kolmogorov equation, there is an SDE which is time-invariant and is called the average SDE. The analysis of the original SDE (12) can approximately be transformed to the average SDE. Therefore, the problem in the switching case in this section can be reduced to the one in fixed case in last section.

**Backward Kolmogorov equation for SDE (12).** Denote $Z$ as the state of (12); that is, $Z = \text{col}[x^\alpha, \theta^\alpha, \lambda^\alpha, \nu^\alpha]$. The stochastic process $(Z(t), \sigma(t))$, rather than the $Z(t)$, is a Markovian process, whose infinitesimal is $\mathcal{A}^\alpha = \frac{1}{\alpha} Q + \mathcal{A}_s$, where $\mathcal{A}_s$ is the infinitesimal of the $s$-subsystem of (6), $Q$ is the infinitesimal of the Markov chain $\sigma(t)$. This means that the average number of jumps of $\sigma(t)$ per unit of time is proportional to $1/\alpha$. Let $\phi$ be a sufficiently smooth real-valued function defined on the state space $(Z, \sigma)$ and let $W(t, Z, s) = \mathbb{E}[\phi(Z(t), \sigma(t))|Z(0) = Z, \sigma(0) = s]$. From the standard analysis in stochastic theory, $W(t, Z, s)$ is a unique bounded classical solution to the following partial differential equation with the initial data $W(0, Z, s) = \phi(Z, s)$:

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

where $\overrightarrow{W}(t, Z) = \left(W(t, Z, 1), \cdots, W(t, Z, S)\right)^T$ and $Q \overrightarrow{W}(t, Z)[s]$ denotes the $s$-row of the matrix $Q \overrightarrow{W}(t, Z)$. The partial differential Eq. (15) is termed as the backward Kolmogorov equation associated with the SDE (12).

**Average backward Kolmogorov equation for SDE (12).** We single out for (15) an approximate solution of the form $W = W_0 + \alpha W_1 + O(\alpha^2)$. Inserting this expression into (15) and equating coefficients of $\alpha^{-1}$ on both sides yields $Q \overrightarrow{W}_0 = 0$. Due to Assumption 4 or its equivalent characterization of (14), one sees that the null space of the adjoint generator $Q^T$ consists of only constants, which also amounts to saying that the null space of its infinitesimal generator $Q$ consists of only constant functions. This fact, together with $Q \overrightarrow{W}_0 = 0$, implies that $\overrightarrow{W}_0$ is a function independent of the switching mode $s \in \mathcal{S}$; that is, $W_0(t, Z, 1) = W_0(t, Z, 2) = \cdots = W_0(t, Z, S)$. For ease of notation, we denote them by $W_0(t, Z)$. Similarly, inserting the expression of $W$ into (15) and equating coefficients of $\alpha^0$ on both sides yields

$$Q \overrightarrow{W}_1 = \left(\begin{array}{c} \frac{\partial W_0}{\partial t} - A_1 W_0 \\ \vdots \\ \frac{\partial W_0}{\partial t} - A_S W_0 \end{array}\right).$$

Recall the Fredholm alternative (see, e.g., [8, pp.641 Theorem 5(iii)] for general case, but only a simple version in finite dimension is needed here) which deals with the solvability of an inhomogeneous linear algebraic equation $Ax = b$ with $A \in \mathbb{R}^{n \times n}$ and $b \in \mathbb{R}^n$; that is, this inhomogeneous equation is solvable if and only if $b$ belongs to the column space of $A$, which is the orthogonal complement of $\text{ker}(A^T)$. A direct application of the Fredholm alternative to the Eq. (16) tells us that $(\frac{\partial W_0}{\partial t} - A_1 W_0, \cdots, \frac{\partial W_0}{\partial t} - A_S W_0)^T$ is perpendicular to the null space of $Q^T$. Noting that $\text{Null}(Q^T) = \pi$ in (14), it is nature to have $\sum_{s=1}^S \pi_s \left(\frac{\partial W_0}{\partial t} - A_s W_0\right) = 0$, which gives $\sum_{s=1}^S \pi_s W_0(t, Z, s) = 0$. Therefore, $W_0(t, Z) = 0$. This fact and (15) together imply that $W(t, Z, s) = \phi(Z, s)$.
rise to an average backward equation

\[
\frac{\partial W_0}{\partial t} = A_\pi \cdot W_0,
\]

(17)

where \( A_\pi = \sum_{s=1}^{S} \pi_s A_s \) is the stochastic average of the infinitesimal generators \( A_s \) with respect to the invariant measure \( \pi \). As an conclusion, we have shown that the first term \( W_0 \) in the series of \( W \) is a solution to another backward equation specified by the average operator \( A_\pi \).

Similar equations as (16) and (17) have been reported in [26]; however, the analysis presented here holds for SDE, while [26] works only for ordinary differential equations. Therefore, the averaging method is extended from deterministic case in [26] to stochastic case in the current paper. While [28] also considers the stochastic averaging, theoretical justifications such as the analysis of the backward Kolmogorov equation and average backward Kolmogorov equation are not provided there.

**Average SDE for** (12). We proceed to construct for the average backward Eq. (17) an SDE whose infinitesimal is exactly \( A_\pi \). We call such an SDE as the average equation for (12). Denoting the drift vector for the \( s \)-subsystem in Eq. (12) by \( a_s \) and the diffusion matrix (the diffusion matrix for an SDE \( dx = f dt + g dw \) is defined as \( gg^T \)) by \( \Gamma_s \), and noting \( A_s = a_s \cdot \nabla + \frac{1}{2} \Gamma_s : \nabla \nabla \), the average operator \( A_\pi \) can be calculated as \( A_\pi = \pi_\pi \cdot \nabla + \frac{1}{2} \Gamma_\pi : \nabla \nabla \), where \( \pi_\pi = \sum_{s=1}^{S} \pi_s a_s \) and \( \Gamma_\pi = \sum_{s=1}^{S} \pi_s \Gamma_s \) are, respectively, the averages of the drift vector and the diffusion matrix. Corresponding to \( A_\pi \) above, one can construct an average SDE for (12) as follows

\[
\begin{align*}
\dot{x} &= -c \mathcal{L}_\pi \cdot \bar{x} - \dot{\theta} - \nabla F(\bar{x}) - \bar{\lambda} \odot \nabla G(\bar{x}) - \bar{v} \odot \nabla H(\bar{x}) \, dt + c \tilde{M} d\bar{w}(t), \\
\dot{\theta} &= c \mathcal{L}_\pi \cdot \bar{x}, \\
\dot{\lambda} &= \left[ \bar{\lambda} \odot (1 + \eta \odot \bar{\lambda}) \right] \odot G(\bar{x}), \\
\dot{v} &= H(\bar{x}),
\end{align*}
\]

(18)

where \( \mathcal{L}_\pi = \sum_{s=1}^{S} \pi_s \mathcal{L}_s \) is the stochastic average of the Laplacians \( \{ \mathcal{L}_s, s \in S \} \) for the graphs \( \{ G_s, s \in S \} \), \( \tilde{M} \in \mathbb{R}^{(nN) \times N^2} \) is chosen such that \( \tilde{M} \tilde{M}^T = \sum_{s=1}^{S} \pi_s \mathcal{M}_s \mathcal{M}_s^T \in \mathbb{R}^{(nN) \times (nN)} \).

As we have mentioned before and will show it later, the average system (18) acts as the role of approximating the original optimization dynamical system (12). Therefore, the stability of the optimization dynamical systems (12) can be inferred from that of the average system (18). With this consideration, we first show the stability of the average system (18). Due to Assumption 4 and similar as the proof of Lemma 3.4 in [26], the average graph Laplacian \( \mathcal{L}_\pi \) can be shown to have a simple zero eigenvalue, and thus it can be viewed as a Laplacian for a certain fixed connected graph. Replacing \( \mathcal{L} \) and \( \mathcal{M} \) in Theorem 5.1 with \( \mathcal{L}_\pi \) and \( \tilde{M} \), respectively, modifying Assumption 3 into Assumption 3’, and arguing in a similar line of the proof for Theorem 5.1, we establish a stability result for the average system (18) in the following theorem.

**Theorem 6.1** Consider the average system (18) for the constrained optimization problem (1) under Assumptions 1, 2, 3’, and 4 with \( \kappa \leq \sqrt{\bar{\lambda}_2}/2 \). Then for any initial
conditions with \( \lambda_{ij}(0) > 0 \) and \( \sum_{i=1}^{N} \theta_{i}(0) = 0 \), one has \( \lim_{t \to \infty} \|\bar{x}_i(t) - x^*\| = 0 \) almost surely, where \( j = 1, \ldots, r_i, i = 1, \ldots, N \).

With the stability of the average system, we proceed to prove the stability of the original optimization dynamics (12) by exploring a connection that the dynamics (12) can be approximated by the average system (18). Associate with the systems (12) and (18), there are, respectively a backward Kolmogorov equation (15) and an average backward Eq. (17), whose solutions are, respectively, \( \delta \) and (18), there are, respectively a backward Kolmogorov equation (15) and an average backward Eq. (17), whose solutions are, respectively, \( W(t, Z, s) \) and \( W_0(t, \bar{Z}) \).

Proof

Consider the distributed optimization algorithm (5) for the constrained optimization problem (1) under Assumptions 1, 2, 3’ and 4 with \( \kappa \leq \sqrt{\lambda_2}/2 \). Then, with a sufficiently small \( \alpha > 0 \) and for any initial conditions with \( \lambda_{ij}(0) > 0 \) and \( \sum_{i=1}^{N} \theta_{i}(0) = 0 \), one has \( \lim_{t \to \infty} \|x_i(t) - x^*\| = 0 \) almost surely, where \( j = 1, \ldots, r_i, i = 1, \ldots, N \).

Proof

The proof amounts to showing the almost sure stability of the system (12). To this end, also consider the Lyapunov candidate \( V \) as in Theorem 5.1 (c.f. Appendix A2). The value of the function \( V \) on the trajectory of (12) is denoted as \( V(t) \). Also define \( V^{[s]}(t) = V(t) 1_{[\sigma(t)/\alpha]\in s} \).

Obviously, \( V(t) = \sum_{s=1}^{S} V^{[s]}(t) \) almost surely. Denote \( d \) the differential of \( V \) along the trajectory of (12). Therefore,

\[
\mathbb{E}[dV] = \sum_{s=1}^{S} \mathbb{E}[dV^{[s]}] = \sum_{s=1}^{S} \mathbb{E}[d(V \cdot 1_{[\sigma(t)/\alpha]\in s})]
\]

\[
= \sum_{s=1}^{S} \mathbb{E}[dV \cdot 1_{[\sigma(t)/\alpha]\in s}] + \sum_{s=1}^{S} \mathbb{E}[V \cdot d(1_{[\sigma(t)/\alpha]\in s})]
\]

\[
= \sum_{s=1}^{S} \mathbb{E}[A_{s} V \cdot d(1_{[\sigma(t)/\alpha]\in s})]dt + \sum_{s=1}^{S} \mathbb{E}[V \cdot d(1_{[\sigma(t)/\alpha]\in s})]
\]

By Lemma 6.1, the above \( \mathbb{E}[dV \cdot 1_{[\sigma(t)/\alpha]\in s}] \) can be calculated as \( \mathbb{E}[V(t) \cdot d(1_{[\sigma(t)/\alpha]\in s})] = \frac{1}{\alpha} \sum_{j=1}^{S} q_{js} \mathbb{E}[V^{[j]}(t)]dt + o(dt) \). Therefore,

\[
\mathbb{E}[dV] = \mathbb{E}[V dt] + (1/\alpha) \sum_{s=1}^{S} \sum_{j=1}^{S} q_{js} \mathbb{E}[V^{[j]}(t)]dt + o(dt)
\]

\[
= \mathbb{E}[V dt] + (1/\alpha) \sum_{s=1}^{S} \sum_{j=1}^{S} q_{js} \mathbb{E}[V^{[j]}(t)]dt + o(dt)
\]

\[
= \mathbb{E}[V dt] + o(dt),
\]

where the last equality uses \( \sum_{s=1}^{S} q_{js} = 0 \) for \( j \in \{1, \ldots, S\} \).

We now calculate \( \mathbb{E}[V dt] \) by arguing in an entirely similar manner as in the proof of Theorem 5.1, with only a modification of the calculation of \( \mathbb{E}[V_{1}] \) in (19) by replacing
\( \mathcal{L} \) in (19)–(20) with \( \mathcal{L}_{\pi} \). The resulting result on \( \mathcal{A}_{\pi} V \), similar as the one in (10), can be calculated as

\[
\mathcal{A}_{\pi} V \leq \Phi_{\pi}(x^*, \theta, \lambda, v) - \Phi_{\pi}(x^*, \theta^*, \lambda^*, v^*) - (\tilde{\lambda}_{2} - 1)(x - x^*)^T (x - x^*)
\]

where \( \Phi_{\pi}(x^*, \theta, \lambda, v) \) is similarly defined as \( \Phi(x^*, \theta, \lambda, v) \) by replacing \( \mathcal{L} \) with \( \mathcal{L}_1 + \cdots + \mathcal{L}_s \) and \( h \) (cf. (9)) with \( h_{\pi} = \frac{1}{2}(c_{\pi_{\min}} - \frac{3}{2}c^2 k^2 \pi_{\max}) \). The rest part of proving asymptotic stability almost surely is similar to the third part of proof for Theorem 5.1.

\( \square \)

The convergence analysis of our algorithm is relatively simple in several aspects in comparison with some existing results. Firstly, when dealing with the optimization constraints, we depart from the use of projected gradient flow whose stability analysis depends heavily on the sophisticated theory of non-smooth analysis and the complex tool of variational inequalities; instead we resort to the elegant Lyapunov argument to prove stability. Secondly, when dealing with switching topologies, we do not build our analysis on the main-stream product theory of stochastic matrices which is hard to generalize to include optimization constraints or communication noises since otherwise the resulting matrices are not stochastic matrices. The multi-scale approach proposed in our paper does not have this limitation since it reduces the optimization algorithm in stochastic networks to that in fixed network and consequently entails a relatively easy convergence analysis.

7 Simulation

Consider the optimization problem (1) on a network with 5 agents. The five local cost functions for five agents are given as

\[
\begin{align*}
    f_1(x_1, x_2) &= 4x_1^2 + 2x_2, \\
    f_2(x_1, x_2) &= 2x_2^2, \\
    f_3(x_1, x_2) &= 4x_1x_2, \\
    f_4(x_1, x_2) &= 2x_2, \\
    f_5(x_1, x_2) &= e^{3x_1+x_2}.
\end{align*}
\]

Some agents may have constraints and while others are free of constraints if corresponding constraint functions are set to zero. For simulation, we choose the case that agent 1 has both inequality and equality constraints with constraint functions

\[
\begin{align*}
    g_1(x_1, x_2) &= (x_1 - 2)^3 - x_2 + 1, \\
    h_1(x_1, x_2) &= 2x_1 - x_2, \\
    g_2(x_1, x_2) &= -x_2 + 2.
\end{align*}
\]

It can be checked that all these functions are convex and the constrained set is nonempty. The true optimal solution and optimal value for this problem are

\[
(x_1^*, x_2^*) = (1, 2) \quad \text{and} \quad \tilde{f}(x_1^*, x_2^*) = 172.41,
\]

respectively.

We now use the DCO algorithm (5) to help check the results. Let \( x_i \in \mathbb{R}^2 \) be the state of agent \( i \in \{1, \cdots, 5\} \), and its dynamics obeys the algorithm (5).

Referring to Fig. 1, the coupling of five agents forms a network which is modeled by a stochastically switching among six possible undirected graphs \( \{G_1, \cdots, G_6\} \) (up of Fig. 1), with the switching rule described by a continuous-time Markov chain

\[
\sigma : [0, +\infty) \to \{1, \cdots, 6\}
\]

with infinitesimal generator

\[
\mathcal{G}.
\]

Springer
These results satisfy the KKT conditions in the Eq. (2) and verify Eq. (4). Due to space limitation, the time evolutions of the states for $\theta$, $\lambda$, $\nu$ are not plotted here. For

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

Obviously, this Markov chain is ergodic and the invariant measure can be calculated as $\pi = (0.1443, 0.2000, 0.1882, 0.1652, 0.1132, 0.1891)$, and whose sample path of $\sigma$ is shown in bottom of Fig. 1. Note that all graphs are very "sparsely connected," implying that less communication resources are required at each time. This advantage is more obvious when the number of agents is large. In this sense, switching networks can save communication resources.

For simulation, to satisfy the condition $\bar{h}\lambda_2^G - 1 > 0$ in Theorem 5.1 or $\bar{h}\lambda_2 - 1 > 0$ in Theorem 6.2, we chose the noise intensities $\sigma_{ij} = 0.1$ and the coupling strength $c = 2$. The parameter $\eta$ is only required to be positive so that we chose $\eta = 1$ for brevity. The initial states of five agents are $x_1(0) = (-2, 4)^T$, $x_2(0) = (-3, 3)^T$, $x_3(0) = (1, -2)^T$, $x_4(0) = (4, 2)^T$, $x_5(0) = (-3, -4)^T$, $\theta_1(0) = \cdots = \theta_5(0) = 1$, $\lambda_1(0) = 3$, $\lambda_2(0) = 3$, $\nu(0) = 3$. The time evolution of the $x$-states for five agents are illustrated in Fig. 2, where the first component of each state $x_i$ asymptotically converges to 1 almost surely (subfigure (a)) and the second component of each state $x_i$ asymptotically converges to 2 almost surely (subfigure (b)). Here, the almost sure convergence means that there exists a set $\mathcal{I}$ of initial states with zero probability measure, such that any trajectory originating from the outside of $\mathcal{I}$ would converge. Therefore, each state $x_i$ of the five agents converges to the optimal solution $(1, 2)$ almost surely. The multipliers $\lambda_1$ and $\lambda_2$ corresponding to the inequality constraints $g_1 \leq 0$ and $g_2 \leq 0$ converge to 0 and 395.0329, respectively. The multiplier $\nu$ corresponding to the equality constraint $h_1 = 0$ converges to $-230.619$. The asymptotic value of the 10-dimensional vector $\theta^*$ is $(453.2396, -232.6197, 0, 387.0329, -8, -4, 0, -2, -445.2395, -148.4132)$. These results satisfy the KKT conditions in the Eq. (2) and verify Eq. (4). Due to space limitation, the time evolutions of the states for $\theta$, $\lambda$, $\nu$ are not plotted here. For

![Image](image_url)
given values of $\sigma_{ij}$, those consistent $c$ with bigger values render convergence of our optimization algorithm to optimal solution with less time.

A quick comparison of our method with projection-dependent methods can be made by analyzing the complexities of the latter. Generally, the projection $P_C(u)$ of a gradient vector $u$ onto a convex set $C$ is difficult to compute since one has to solve a minimum norm problem $\min\{||x - u|| : x \in C\}$. However, the problem becomes much easier if the constrained set $C$ is a half-space. Considering this, a popular way to compute projection is to replace $P_C$ by computing a sequence of projections $P_{C_k}$ onto half-spaces $C_k$ containing the original constrained set. This idea can be found earlier in [12] and was later followed in [42]. As for the constrained set $C = \cap_{i=1}^m C_i = \cap_{i=1}^m \{x \in \mathbb{R}^n | c_i(x) \leq 0\}$ considered in our paper which is a finite level sets of convex functions, the projection of a vector $u$ onto the set $C$ can be computed recursively as $x_{n+1} = \lambda_n u + (1 - \lambda_n) P_{C_m} \cdots P_{C_2} P_{C_1} x_n$, $n = 0, 1, 2, \cdots$, where $x_0$ is the initial value, $\lambda_n$ is in $(0, 1)$ and $C_n$ is a sequence of half-spaces containing $C^i$ for $i = 1, \ldots, m$. Noting that, for projection-dependent methods, each iteration contains a sequence of sub-iterations to compute the gradient projection, the overall complexity of the projection methods is tremendously increased by multiplying that of the projection algorithm. Our method avoids using gradient project and it is hoped that computation complexity can be reduced.

8 Conclusion

This paper has proposed a novel multi-scale method for the distributed convex optimization problem with constraints and presented in a unified framework to address challenging issues like optimization constraints, communication noises and stochastic networks. To overcome the technical obstacle in computing the projection in the pres-
ence of optimization constraints, we design a projection-free, smooth optimization dynamics for easier analysis and simulation. By the stochastic averaging approach in this paper, the dynamics of the algorithm under switching topologies can be approximated by an average system which corresponds to an optimization algorithm under a fixed topology. Therefore, provided the distributed optimization under fixed topology can compute the optimal solution, its extension to switching topologies can also compute the optimal solution. As a consequence, a major advantage of the proposed multi-scale method presented in this paper is that it generalizes previous distributed convex optimization algorithms from a fixed network topology to the switching case. Also, the stochastic averaging in this paper is a generalization of the deterministic averaging in our earlier works, again thanks to the multi-scale method used in this paper.

Acknowledgements The first author would like to thank Professor Zhong-Ping Jiang for his discussions on this paper when he visited New York University. This work is supported by the NSFC of China under the Grants 61663026, 61963028, 62066026, 61866023 and Jiangxi NSF Under Grant 20192BAB207025.

Appendix A: Derivation of The SDE (6)

To characterize the noisy term $\omega_i := \sum_{j \in Mathcal{M}_i(t)} \sigma_{ji} \xi_j (x_j - x_i)$ in Eqs. (5a)–(5b), define $\xi_i = [\xi_{1i}, \xi_{2i}, \ldots, \xi_{Ni}]^T, i = 1, \ldots, N$. Also, for each switching mode $s \in Mathcal{S}$, define $\sigma_{i}^{s} = [a_i^{s} \sigma_{1i}(x_1 - x_i), \ldots, a_i^{sN} \sigma_{Ni}(x_N - x_i)].$ Then the noisy term $\omega_i$ above can be written as $\omega_i = Mathcal{M}_i^s \xi_i.$ Therefore, Eq. (5a) becomes

$$dx_i = \left[ c \sum_{j \in Mathcal{M}_i(t)} (x_j - x_i) - \theta_i - \nabla f_i(x_i) - \sum_{j} \lambda_{ij} \nabla g_{ij}(x_i) - \sum_{i} \nu_{ij} \nabla g_{ij}(x_i) \right] dt + c Mathcal{M}_i^s dw_i,$$

where $\xi_i dt = dw_i$ with $w_i$ an $N$-dimensional standard Brownian motion on the probability space $(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\}_{t \geq 0})$. Let $x = \text{col}(x_1, \ldots, x_N).$ Define the stacked functions $G(x) = \text{col}(\text{col}(g_{ij}(x_1)), \ldots, g_{ij}(x_N)) \in \mathbb{R}^r$, $H(x) = \text{col}(\text{col}(h_{ij}(x_1)), \ldots, h_{ij}(x_N)) \in \mathbb{R}^s$ and the stacked gradients $\nabla F(x) = \text{col}(\text{col}(\nabla f_i(x_1)), \ldots, \nabla f_i(x_N)) \in \mathbb{R}^{nN}$, $\nabla G(x) = \text{col}(\text{col}(\nabla g_{ij}(x_1)), \ldots, \nabla g_{ij}(x_N)) \in \mathbb{R}^{nN}$, $\nabla H(x) = \text{col}(\text{col}(\nabla h_{ij}(x_1)), \ldots, \nabla h_{ij}(x_N)) \in \mathbb{R}^{sn}.$

For each switching mode $s \in Mathcal{S}$, set $\mathcal{M}_s = \text{diag}(\mathcal{M}_i^s) \in \mathbb{R}^{nN \times N^2}$. In addition, let $w = \text{col}(w_i) \in \mathbb{R}^{nN}$ and define $r = \sum_{i=1}^{N} r_i, s = \sum_{i=1}^{N} s_i.$ Set $\lambda = \text{col}(\lambda_i, \ldots, \lambda_N) \in \mathbb{R}^{r}$ with $\lambda_i = \text{col}(\lambda_{i1}, \ldots, \lambda_{in_i}) \in \mathbb{R}^{r_i}$, and $\nu = \text{col}(\nu_1, \ldots, \nu_N) \in \mathbb{R}^{s}$ with $\nu_i = \text{col}(\nu_{i1}, \ldots, \nu_{in_i}) \in \mathbb{R}^{s_i}$. Define $\theta = \text{col}(\theta_1, \ldots, \theta_N) \in \mathbb{R}^{nN}$.

With these, Eq. (5) can be written in a compact form as in (6).

Appendix B: Proof of The Inequality (10)

Firstly, for $V_1$, defining $\chi = \nabla F(x) + \lambda \odot \nabla G(x) + \nu \odot \nabla H(x)$ and $\chi^* = \nabla F(x^*) + \lambda^* \odot \nabla G(x^*) + \nu^* \odot \nabla H(x^*)$ and noting $\theta^* = -\chi^*$ in view of (7), the action of the infinitesimal operator $A$ on $V_1$ can be calculated as

Springer
\[ AV_1 = (x - x^*)^T \left[-cLx - \theta - \nabla F(x) + \frac{3}{2} c^2 tr(\mathcal{M}^T \mathcal{M}) \right] \]

\[ - \sum_{i=1}^{N} \sum_{j=1}^{r_i} \lambda_{ij}(x_i - x^*)^T \nabla g_{ij}(x_i) + [(x - x^*) + (\theta - \theta^*)]^T \left[-(\theta - \theta^*) - (x - x^*) \right] \]

\[ - \sum_{i=1}^{N} s_i v_{ij}(x_i - x^*)^T \nabla h_{ij}(x_i). \]  

(19)

The trace tr(\mathcal{M}^T \mathcal{M}) brings difficulty to the convergence analysis. To get rid of this difficulty, we give an estimation of tr(\mathcal{M}^T \mathcal{M}) as follows,

\[ \text{tr}(\mathcal{M}^T \mathcal{M}) = \sum_{i=1}^{N} \text{tr}((\mathcal{M}^i)^T \mathcal{M}^i) = \sum_{i=1}^{N} \sum_{j=1}^{N} (a_{ij}^2)(\sigma_{ji})^2(x_j - x_i)^T(x_j - x_i) \]

\[ \leq \kappa^2 \sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij}(x_j - x_i)^T(x_j - x_i) = \kappa^2 x^T L x. \]  

(20)

As for Part A in (19), noting \( x^T L x = (x - x^*)^T L (x - x^*) \) and \( h = \frac{1}{2}(c - \frac{3}{2} c^2 \kappa^2) \) (c.f. Eq. (9)), the part A can be estimated as \( A \leq (x - x^*)^T[-hLx - \theta - \nabla F(x)] - h(x - x^*)^T L (x - x^*). \) Noting that the square bracket is exactly the minus gradient of \( \Psi(x, \theta) \) (c.f. Eq. (9)) with respect to \( x \) and denoting by \( \lambda_{\theta}^G \) the smallest nonzero eigenvalue of the Laplacian \( L \), one obtains \( A \leq (x - x^*)^T \nabla_x \Psi(x, \theta) - h\lambda_{\theta}^G (x - x^*)^T (x - x^*) \leq \Psi(x^*, \theta) - \Psi(x, \theta) - h\lambda_{\theta}^G (x - x^*)^T (x - x^*) \), where the last inequality uses the fact that \( \Psi(x, \theta) \) is convex in its first argument. As for part B in (19), we can prove the inequality \( B \leq (x - x^*)^T (\theta - \theta^*) + (x - x^*)^T (x - x^*) \) by rewriting it into a quadratic form in terms of \( x - x^*, \theta - \theta^*, x - x^* \) and by showing the corresponding matrix to be semi-positive definite. In view of \( (x - x^*)^T (\theta - \theta^*) = \Psi(x, \theta) - \Psi(x, \theta^*) \), one obtains \( A + B \leq [\Psi(x^*, \theta) - \Psi(x, \theta^*)] - (h\lambda_{\theta}^G - 1)(x - x^*)^T (x - x^*) \). Now,

\[ AV_1 \leq \Psi(x^*, \theta) - \Psi(x, \theta^*) - (h\lambda_{\theta}^G - 1)(x - x^*)^T (x - x^*) + \sum_{i=1}^{N} \sum_{j=1}^{r_i} \lambda_{ij} g_{ij}(x_i) \]

\[ - \sum_{i=1}^{N} \sum_{j=1}^{r_i} \lambda_{ij} g_{ij}(x_i) + \sum_{i=1}^{N} \sum_{j=1}^{s_i} v_{ij} h_{ij}(x^*) - \sum_{i=1}^{N} \sum_{j=1}^{s_i} v_{ij} h_{ij}(x_i), \]

where we use the convexity of the functions \( g_{ij} \) and \( h_{ij} \).

Secondly, we calculate the action of the infinitesimal operator \( \mathcal{A} \) on \( V_2 + V_3 \). Firstly note that the function \( V_3 \) can be calculated by the definition of Bregman divergence
as \( V_3 = \sum_{i=1}^{N} \sum_{j=1}^{r_i} (\lambda_{ij} - \lambda_{ij}^*) - \sum_{(i,j) \in \Omega} \lambda_{ij}^* (\ln \lambda_{ij} - \ln \lambda_{ij}^*) \). Therefore,

\[
\mathcal{A} V_2 + \mathcal{A} V_3 = \sum_{i=1}^{N} \sum_{j=1}^{r_i} \eta_{ij} \lambda_{ij} (\lambda_{ij} - \lambda_{ij}^*) \frac{g_{ij}(x_i)}{1 + \eta_{ij} \lambda_{ij}} + \sum_{i=1}^{N} \sum_{j=1}^{r_i} \lambda_{ij} \frac{g_{ij}(x_i)}{1 + \eta_{ij} \lambda_{ij}} - \sum_{(i,j) \in \Omega} \lambda_{ij}^* g_{ij}(x_i).
\]

Furthermore, the action of the infinitesimal operator \( \mathcal{A} \) on \( V_4 \) can be easily calculated as \( \mathcal{A} V_4 = \sum_{i=1}^{N} \sum_{j=1}^{r_i} (v_{ij} - v_{ij}^*) h_{ij}(x_i) \). Collecting above results for \( \mathcal{A} V_i, i = 1, 2, 3, 4 \) and recalling the definition of \( \Phi \) in (8), one obtains the inequality (10).

**Appendix C: Proof of The Saddle Point Conditions (11)**

Due to convexity and affinity, the following four results hold: \( \Sigma_{i=1}^{N} f_i(x_i) + h x L x \geq \Sigma_{i=1}^{N} f_i(x^*) + \Sigma_{i=1}^{N} \nabla f_i(x^*) (x_i - x^*), g_{ij}(x_i) \geq g_{ij}(x^*) + \nabla g_{ij}(x^*) (x_i - x^*), h_{ij}(x_i) = h_{ij}(x^*) + \nabla h_{ij}(x^*) (x_i - x^*), \) and \( x - x^* = \bar{X}^* - x^* \). Multiplying them by \( 1, \lambda_{ij}^*, v_{ij}^*, \theta^* \), respectively (for the forth equality we use \( (x - x^*)^T \theta^* \) for multiplication), and adding gives \( \Phi(x, \theta^*, \lambda^*, v^*) \geq \Phi(x^*, \theta^*, \lambda^*, v^*) + [\theta^* + \nabla F(x^*) + \lambda^* \nabla G(x^*) + v^* \nabla H(x^*)]^T (x - x^*) \). In view of Eq. (7), the terms in the square bracket sum to be zero. Therefore, \( \Phi(x, \theta^*, \lambda^*, v^*) \geq \Phi(x^*, \theta^*, \lambda^*, v^*) \).

On other hand, noting that \( \lambda_{ij} g_{ij}(x_i^*) \leq 0 = \lambda_{ij}^* g_{ij}(x_i^*) \) and \( h_{ij}(x^*) = 0 \), it can be directly checked that \( \Phi(x^*, \theta, \lambda, v) \leq \Phi(x^*, \theta^*, \lambda^*, v^*) \). This inequality, together with the inequality derived in last paragraph, gives rise to the saddle point condition (11).

**Appendix D: Proof of “\( \mathcal{A} V = 0 \Rightarrow (x, \theta, \lambda, v) = (x^*, \theta^*, \lambda^*, v^*) \)”**

Letting \( \mathcal{A} V = 0 \) gives \( (x - x^*)^T (x - x^*) = 0 \) and \( \Phi(x^*, \theta, \lambda, v) = \Phi(x^*, \theta^*, \lambda^*, v^*) \), where the former implies \( x = x^* \) and the latter, together with the fact that \( (x^*, \theta^*, \lambda^*, v^*) \) is a saddle point of the Lagrangian \( \Phi \), implies \( \Phi(x^*, \theta, \lambda, v) = \Phi(x, \theta^*, \lambda^*, v^*) = \Phi(x^*, \theta^*, \lambda^*, v^*) \). Recall that \( x^* = \text{1}_N \otimes x^* \) with \( x^* \) being the optimal solution which satisfies \( g_{ij}(x^*) \leq 0 \) in the KKT condition (2a). Inserting \( x_i = x^* \) into (5c) yields \( \dot{\lambda}_{ij} = \frac{\lambda_{ij} g_{ij}(x^*)}{1 + \eta_{ij} \lambda_{ij}} \). If \( g_{ij}(x^*) = 0 \), then \( \dot{\lambda}_{ij} = 0 \) which implies that \( \lambda_{ij}(t) \) stays positive for all \( t \geq 0 \) since the initial value of \( \lambda_{ij} \) is chosen to be positive. If \( g_{ij}(x^*) < 0 \), then with \( g_{ij}(x^*) = -a < 0 \), one has \( \dot{\lambda}_{ij} = \frac{-a}{1 + \eta_{ij} \lambda_{ij}} \), whose trajectory can be shown by elementary analysis as \( \lambda_{ij}(t) \geq 0 \) for \( t \geq 0 \) since the initial value of \( \lambda_{ij} \) is chosen to be positive. In short, in both cases, for Eq. (5c) with \( x_i = x^* \), one has that \( \lambda_{ij}(t) \geq 0, \forall t \geq 0 \) provided the initial value is positive, and consequently \( \dot{\lambda}_{ij}(t) \leq 0 \) since \( g_{ij}(x^*) \leq 0 \). Therefore, \( \lambda_{ij} \geq \lambda_{ij}^* \). On the other hand, the fact \( \Phi(x^*, \theta, \lambda, v) = \Phi(x^*, \theta^*, \lambda^*, v^*) \) yields \( \sum_{i=1}^{N} \sum_{j=1}^{r_i} (\lambda_{ij} - \lambda_{ij}^*) g_{ij}(x^*) = 0 \). Since \( g_{ij}(x^*) \leq 0 \) and \( \lambda_{ij} \geq \lambda_{ij}^* \), each sum in this summation is non-positive and therefore

\[ \square \] Springer
(\lambda_{ij} - \lambda_{ij}^*) g_{ij}(x^*) = 0$, i.e., $\lambda_{ij} g_{ij}(x^*) = \lambda_{ij}^* g_{ij}(x^*) = 0$. This fact implies that the right-hand side of Eq. (5c) with $x_i = x^*$ is zero and thus $\lambda_{ij} = \lambda_{ij}^*$ for some constant $\lambda_{ij}^* \geq 0$. Thus, $\lambda_{ij}^* g_{ij}(x^*) = 0$. Inserting $x^* = 1_N \otimes x^*$ into Eqs. (5b) and (5d) gives that $\theta_i = \theta_i^*$ and $v_{ij} = v_{ij}^*$ for some constants $\theta_i^*$ and $v_{ij}^*$. Inserting $(x^*, \theta_i^*, \lambda_{ij}^*, v_{ij}^*)$ into (5a) gives $\sum_{i=1}^N \nabla f_i(x^*) + \sum_{i=1}^N \sum_{j=1}^{r_i} \lambda_{ij}^* \nabla g_{ij}(x^*) + \sum_{i=1}^N \sum_{j=1}^{s_i} v_{ij}^* \nabla h_{ij}(x^*) = 0$. In conclusion, the KKT conditions (2) are satisfied at the point $(x^*, \lambda^*, v^*)$. Then by uniqueness of multipliers in (4), $\lambda_{ij}^* = \lambda_{ij}^*$, $v_{ij}^* = v_{ij}^*$.

Also, by differentiating both sides of $\Phi(x, \theta^*, \lambda^*, v^*) = \Phi(x^*, \theta^*, \lambda^*, v^*)$ with respect to $x$ and then by enforcing $x = x^*$, one obtains $\theta + \nabla F(x^*) + \lambda^* \otimes \nabla G(x^*) + v^* \otimes \nabla H(x^*) = 0$. This, combined with Eq. (7), gives $\theta = \theta^*$. In conclusion, letting $AV = 0$ gives rise to $(x, \theta, \lambda, v) = (x^*, \theta^*, \lambda^*, v^*)$.

References

1. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via alternating direction method of multipliers. Found. Trends Mach. Learn. 3(1), 1–122 (2011)
2. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
3. Brunner, F.D., Dürr, H.B., Ebenbauer, C.: Feedback design for multi-agent systems: a saddle point approach. In: The 51st IEEE Conference on Decision and Control, pp. 3783–3789, (2012)
4. Carli, R., Fagnani, F., Speranzon, A., Zampieri, S.: Communication constraints in the average consensus problem. IEEE Trans. Autom. Control 44(3), 671–684 (2008)
5. Charalamous, C.: Distributed constrained optimization by consensus-based primal-dual perturbation method. IEEE Trans. Autom. Control 59(6), 1524–1538 (2014)
6. Chen, G., Yi, P., Hong, Y.: Distributed optimization with projection-free dynamics (2021). arXiv:2105.02450
7. Duchi, J.C., Agarwal, A., Wainwright, M.M.: Dual averaging for distributed optimization: convergence analysis and network scaling. IEEE Trans. Autom. Control 57(3), 592–606 (2012)
8. Evans, L.C.: Partial Differential Equations, 2nd edn. American Mathematical Society, Providence (2010)
9. Feijer, D., Paganini, F.: Stability of primal-dual gradient dynamics and applications to network optimization. Automatica 46(12), 1974–1981 (2010)
10. Fragoso, M.D., Costa, O.L.V.: A unified approach for stochastic and mean square stability of continuous-time linear systems with markovian jumping parameters and additive disturbances. SIAM J. Control Optim. 44(4), 1165–1190 (2015)
11. Freedman, D.: Markov Chains. Springer, New York (1983)
12. Fukushima, M.: A relaxed projection method for variational inequalities. Math. Program. 35(1), 58–79 (1986)
13. Haken, H.: Synergetik. Springer, New York (1982)
14. Hosseini, S., Chapman, A., Mesbahi, M.: Online distributed convex optimization on dynamic networks. IEEE Trans. Autom. Control 61(11), 3545–3550 (2016)
15. Jakovetic, D., Xavier, J., Moura, J.M.F.: Fast distributed gradient methods. IEEE Trans. Autom. Control 59(5), 1131–1146 (2014)
16. Kia, S.S., Corts, J., Martnez, S.: Distributed convex optimization via continuous-time coordination algorithms with discrete-time communication. Automatica 55, 254–264 (2015)
17. Li, T., Wu, F., Zhang, J.: Multi-agent consensus with relative-state-dependent measurement noises. IEEE Trans. Autom. Control 59(9), 2463–2468 (2014)
18. Lobel, I., Ozdaglar, A.: Distributed subgradient methods for convex optimization over random networks. IEEE Trans. Autom. Control 56(6), 1291–1306 (2011)
19. Lou, Y., Hong, Y., Wang, S.: Distributed continuous-time approximate projection protocols for shortest distance optimization problems. Automatica 69, 289–297 (2016)
20. Mao, X.: Stochastic versions of the lasalle theorem. J. Differ. Equ. 153, 175–195 (1999)
21. Nedic, A., Ozdaglar, A.: Distributed subgradient methods for multi-agent optimization. IEEE Trans. Autom. Control 54(1), 48–61 (2009)
22. Nedic, A., Ozdaglar, A.: Distributed optimization over time-varying directed graphs. IEEE Trans. Autom. Control 60(3), 601–615 (2015)
23. Nedic, A., Ozdaglar, A., Parrilo, P.A.: Constrained consensus and optimization in multi-agent networks. IEEE Trans. Autom. Control 55(4), 922–938 (2010)
24. Nemirovsky, A.S., Yudin, D.B.: Problem Complexity and Method Efficiency in Optimization. Wiley, Chichester (1983)
25. Ni, W., Wang, X.: Averaging method to distributed convex optimization for continuous-time multi-agent systems. Kybernetika 52(6), 898–913 (2016)
26. Ni, W., Wang, X., Xiong, C.: Leader-following consensus of multiple linear systems under switching topologies: an averaging method. Kybernetika 48(6), 1194–1210 (2012)
27. Ni, W., Wang, X., Xiong, C.: Consensus controllability, observability and robust design for leader-following linear multi-agent systems. Automatica 49(7), 2199–2205 (2013)
28. Ni, W., Zhao, D., Ni, Y., Wang, X.: Stochastic averaging approach to leader-following consensus of linear multi-agent systems. J. Frankl. Inst. 353(12), 2650–2669 (2016)
29. Parikh, N.: Proximal algorithms. Found. Trends Optim. 1(3), 123–231 (2013)
30. Pavliotis, G.V., Stuart, A.M.: Multiscale Methods: Averaging and Homogenization. Springer, New York (2008)
31. Qu, G., Li, N.: Harnessing smoothness to accelerate distributed optimization. IEEE Trans. Control Netw. Syst. 5(3), 1245–1260 (2017)
32. Raginsky, M., Bouvrie, J.: Continuous-time stochastic mirror descent on a network: variance reduction, consensus, convergence. In: The 51st IEEE Conference on Decision and Control, pp. 6793–6800, (2012)
33. Ram, S.S., Nedic, A., Veeravalli, V.V.: Distributed stochastic subgradient projection algorithms for convex optimization. J. Optim. Theory Appl. 147, 516–545 (2010)
34. Ram, S.S., Nedic, A., Veeravalli, V.V.: Incremental stochastic subgradient algorithms for convex optimization. SIAM J. Optim. 20(2), 691–717 (2010)
35. Tsitsiklis, J.N., Bertsekas, D.P., Athans, M.: Distributed asynchronous deterministic and stochastic gradient optimization algorithms. IEEE Trans. Autom. Control 31(9), 803–812 (1986)
36. Wachsmuth, G.: On LICQ and the uniqueness of Lagrange multipliers. Oper. Res. Lett. 41(1), 78–80 (2013)
37. Wang, H., Li, Huaqing H., Zhou, B.: Distributed optimization under inequality constraints and random projections. In: Distributed Optimization, Game and Learning Algorithms, pp. 39–65. Springer, (2021)
38. Wang, J., Elia, N.: Mitigation of complex behavior over networked systems: analysis of spatially invariant structures. Automatica 49(6), 1626–1638 (2013)
39. Wolfowitz, J.: Products of indecomposable, aperiodic, stochastic matrices. Proc. Am. Math. Soc. 14(4), 733–737 (1963)
40. Xie, P., You, K., Tempo, R., Wu, C.: Distributed convex optimization with inequality constraints over time-varying unbalanced digraphs. IEEE Trans. Autom. Control 63(12), 4331–4337 (2018)
41. Xin, R., Khan, U.A.: A linear algorithm for optimization over directed graphs with geometric convergence. IEEE Control Syst. Lett. 2(3), 315–320 (2018)
42. Yang, Q.: The relaxed CQ algorithm solving the split feasibility problem. Inverse Probl. 20(4), 1261–1266 (2004)
43. Yi, P., Hong, Y., Liu, F.: Distributed gradient algorithm for constrained optimization with application to load sharing in power systems. Syst. Control Lett. 83, 45–52 (2015)
44. Yuan, D., Ho, D.W.C., Hong, Y.: On convergence rate of distributed stochastic gradient algorithm for convex optimization with inequality constraints. SIAM J. Control Optim. 54(5), 2872–2892 (2016)
45. Zeng, X., Yi, P., Hong, Y.: Distributed continuous-time algorithm for constrained convex optimizations via nonsmooth analysis approach. IEEE Trans. Autom. Control 62(10), 5227–5233 (2017)
46. Zhu, M., Martinez, S.: On distributed convex optimization under inequality and equality constraints. IEEE Trans. Autom. Control 57(1), 691–719 (2011)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.