Data-driven policy iteration algorithm for optimal control of continuous-time Ito stochastic systems with Markovian jumps

DOI:
10.1049/iet-cta.2015.0973

Accepted and to appear in IET Control Theory and Applications

Data-driven policy iteration algorithm for optimal control of continuous-time Ito stochastic systems with Markovian jumps

Journal:	*IET Control Theory & Applications*
Manuscript ID	CTA-SI-2015-0973.R1
Manuscript Type:	Regular Paper
Date Submitted by the Author:	07-Apr-2016
Complete List of Authors:	Song, Jun; School of Information Science and Engineering, East China University of Science and Technology He, Shuping; Anhui University, School of Electrical Engineering and Automation Liu, Fei; Southern Yangtze Universtiy, Institute of Automation Niu, yugang; East China University of Science & Technology, School of Information Science & Engineering; Ding, Zhengtao; University of Manchester, Control Systems Centre, School of Electrical and Electronic Engineering ;
Keyword:	Adaptive Control, Optimal Control, Stochastic Systems

Data-driven policy iteration algorithm for optimal control of continuous-time Itô stochastic systems with Markovian jumps*

Jun Song[a], Shuping He[b†], Fei Liu[c], Yugang Niu[a], Zhengtao Ding[d]

[a] *Key Laboratory of Advanced Control and Optimization for Chemical Process (East China University of Science and Technology), Ministry of Education, Shanghai 200237, China*

[b] *School of Electrical Engineering and Automation, Anhui University, Hefei 230601, China*

[c] *Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education), Institute of Automation, Jiangnan University, Wuxi 214122, China*

[d] *Control Systems Center, School of Electrical and Electronic Engineering, University of Manchester, Sackville Street Building, Manchester M13 9PL, UK*

Abstract: This paper studies the infinite horizon optimal control problem for a class of continuous-time systems subjected to multiplicative noises and Markovian jumps by using a data-driven policy iteration algorithm. The optimal control problem is equivalent to solve a stochastic coupled algebraic Riccatic equation (CARE). An off-line iteration algorithm is first established to converge the solutions of the stochastic CARE, which is generalized from a implicit iterative algorithm. By applying subsystems transforation (ST) technique, the off-line iterative algorithm is decoupled into N parallel Kleinman's iterative equations. To learn the solution of the stochastic CARE from N decomposed linear subsystems data, a ST-based data-driven policy iteration algorithm is proposed and the convergence is proved. Finally, a numerical example is given to illustrate the effectiveness and applicability of the proposed two iterative algorithms.

Keywords: Data-driven; Markovian jump Itô stochastic systems; policy iteration; decoupling; integral reinforcement learning (IRL); stochastic coupled algebraic Riccati equation

1 Introduction

Stochastic Itô systems and Markovian jump systems have recently received considerable attention, since they may effectively model a class of plants in economics, chemical processes and other areas. As a special case, the stochastic systems with Markovian jumps have also attracted much attention. A variety of results on the analysis and synthesis of such class of systems have been obtained, such as, sliding mode control for stochastic Markovian jump systems [1, 2, 3], exponential filtering for Itô stochastic time-delay systems with Markovian switchings [4],

*This work was supported in part by the NNSF from China (61203051, 61273073), and the Foundation for Distinguished Young Scholars of Anhui Province (1608085J05).

†Corresponding author. Tel: +86 0551 3861413, fax: +86 0551 3861413. Email addresses: shuping.he@ahu.edu.cn(S. He).

globally exponential stabilization for time-delayed stochastic Markovian jump systems [5]. For more details, the readers are referred to [6, 7, 8, 9, 10, 11] and the references therein.

Amongst some of the most important problems in control theory, stability and stabilization in both deterministic setting and stochastic setting have received much more attention than other problems. For deterministic system, it is well known that the stability of a system is equivalent to the existence of a positive definite solution of the associated Lyapunov equation. Similarly, the optimal control of a system is equivalent to the existence of a positive definite solution of the associated Riccatic equation. Till now, many optimal control problems for stochastic systems have been addressed very well by developing numerical iteration algorithms and LMI-based convex optimization approaches. The authors in [12] introduced two parallel algorithms for to solve the continuous-time coupled Lyapunov equations associated with linear Markovian jump systems. The parallel algorithm for discrete-time coupled Lyapunov equations has been investigated in [13]. And, two iterative approaches based on positive operator have been given in [14] for to solve the continuous-time and discrete-time stochastic coupled algebraic Lyapunov equations associated with the stability for stochastic systems with multiplicative noise and Markovian jumps. By using a convex optimization approach, the authors in [15] have studied the H_2/H_∞ control problem for systems with multiplicative noise and Markovian jumps. It should be remarked that the **suboptimal** solutions for the stochastic coupled algebraic Riccatic equation (CARE) was proposed in [15], which was associated with the optimal control problem for Itô stochastic systems with Markovian jumps. Unfortunately, the iterative algorithm for learning the solutions of the stochastic CARE (i.e., ***optimal*** controller) has attracted little research attention, and this constitutes one of the motivations of the present research.

Reinforcement learning (RL) refers to a class of methods that enable the design of adaptive optimal controllers for uncertain dynamical systems. RL method is learning online, in real time, and the solutions to user-prescribed optimal control problems for uncertain systems using only measured data from the controlled system; see [16, 17, 18] and the references therein for more details. Moreover, RL shares some essential features with adaptive dynamic programming (ADP), which is originally inspired by learning mechanisms observed in biological systems [19, 20, 21, 22]. RL concerns how an agent should modify its actions to interact with the unknown environment and to achieve a long-term goal. For the adaptive optimal control problems for continuous/discrete-time linear/nonlinear systems, very many corresponding RL algorithms have been proposed [23, 24, 25, 26, 27, 28, 29, 30, 31]. The data-based policy iteration algorithms in [29] and [32] were developed for the continuous-time H_∞ control problem and linear quadratic regulator problem, respectively. More recently, the studies on RL algorithms for optimal control of stochastic systems also has received increasing attentions. By applying RL methods, a robust ADP algorithm was developed in [33] for the optimal control problem of continuous-time linear stochastic systems with state-dependent noise. The authors in [34] proposed an ADP algorithm for the optimal control problem of unknown discrete-time nonlinear Markov jump systems. For the optimal control problem for continuous-time Markov jump linear systems, the authors in [35] gave a policy iteration algorithm. It is worth to mention that, in order to deal with the strongly coupling relation between the subsystems in Markovian jumping systems, the authors in [35] used a interesting decoupling technique, i.e., Subsystems transformation (ST).

However, to the authors' best knowledge, the relevant researches on the iterative algorithm for converging to the solutions of the stochastic CARE, which associated with optimal control problem for Itô stochastic systems with Markovian jumps, have not fully been investigated in the literature, saying nothing of applying the data-driven policy iteration approaches. In fact, it is difficult to investigate this issue due to the special structure of this kind of systems. Therefore, the problems considered here are of more complexity than those in [35]. Besides, it is easy to see that the existing adaptive optimal control approaches for continuous-time stochastic systems,

such as in [33, 36], also cannot be applied to the stochastic systems with Markovian jumps.

In this work, our aim is to develop an adaptive optimal control algorithm for learning the solutions of the stochastic CARE associated with the optimal control problem for continuous-time Itô stochastic systems with Markovian jumps. Inspired by the implicit iterative algorithm in [14], an off-line iterative algorithm is first proposed for solving the stochastic CARE. Furthermore, by invoking the ST technique and integral reinforcement learning (IRL) approach [37], a data-driven policy iteration algorithm is developed to converge the solutions of the stochastic CARE based on sampling the states of the N decomposed linear subsystems. The two-step policy iteration algorithm can be implemented online via a parallel framework. The convergence proof of this novel algorithm is also been provided. Simulation results demonstrate the effective and validity of the proposed policy iteration algorithm.

The contributions of this paper can be briefly summarized as follows.

(1). An off-line iterative algorithm is presented for the first time for learning the stochastic CARE associated with the optimal control problem for the continuous-time systems subjected to multiplicative noise and Markovian jumps. The iterative algorithm is extended from the implicit iterative algorithm proposed in [14];

(2). The ST technique is introduced to decompose the stochastic systems with Markovian jumps into N decoupled linear subsystems, in which coupling relation is embedded via Q-updating with Kleinman's algorithm [38], and thus the stochastic characteristic is removed;

(3). Collecting the state data from the N reconstructed linear subsystems instead of the original stochastic jump systems, an novel data-driven policy iteration algorithm based on ST technique [35] and IRL approach [37] is developed to learn the solutions of the stochastic CARE and its convergence is proved.

This paper is organized as follows. The problem description and an off-line iteration algorithm are given in Section 2. The ST-based policy iteration algorithm with convergence analysis is proposed in Section 3. A simulation example is provided in Section 4 for demonstrating the effective of the proposed policy iteration algorithm. Finally, a brief conclusion is drawn in Section 5.

2 Backgrounds

In this section, we present some background knowledge. First, we give the problem description of optimal control of continuous-time Itô stochastic systems with Markovian jumps. Second, we briefly review a numerical iteration algorithm, i.e. *implicit iteration algorithm* [14], for to solve the problem.

2.1 Problem description

Throughout this work, let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathcal{P})$ be a given filtered probability space where there exist independent wide sense stationary, second-order processes $w_k(t) \in \mathbf{R}$ $(k \in \{1, r\})$, and a right continuous homogeneous Markov chain $r(t)$, $t \geq 0$ with state space $\mathbf{S} = \{1, 2, \cdots, N\}$. We assume that $r(t)$ is independent of $w_k(t)$ and has the following transition probability:

$$P_r\{r(t+\Delta t) = j | r(t) = i\} = \begin{cases} \pi_{ij}\Delta t + o(\Delta t) & j \neq i, \\ 1 + \pi_{ii}\Delta t + o(\Delta t) & j = i, \end{cases} \tag{1}$$

where $\Delta t > 0$, $\lim_{\Delta t \to 0} \frac{o(\Delta t)}{\Delta t} = 0$, and $\pi_{ij} \geq 0$ $(j \neq i)$ is the transition rate from mode i at time t to mode j at time $t + \Delta t$ and $\pi_{ii} = -\sum_{j \in \mathbf{S}, j \neq i} \pi_{ij}$. $\mathcal{F}_t$ stands for the smallest σ-algebra generated by processes $w_k(s)$, $r(s)$, $0 \leq s \leq t$, i.e. $\mathcal{F}_t = \sigma\{w_k(s), r(s) | 0 \leq s \leq t\}$.

Consider the following continuous-time Itô stochastic system with Markovian jumps

$$\mathrm{d}x(t) = [A_0(r(t))x(t) + B(r(t))u(t)]\,\mathrm{d}t + \sum_{k=1}^{r} A_k(r(t))x(t)\mathrm{d}w_k(t), \tag{2}$$

where $x(t) \in \mathbf{R}^n$ is the state vector, and $u(t) \in \mathbf{R}^m$ is the controlled input; $A_0(r(t))$, $B(r(t))$ and $A_k(r(t))$ are constant matrices with appropriate dimensions.

Subject to system (1)-(2) which is assumed to be stochastically stabilizable [39], our objective is to find a set of mode-dependent control policy $\psi(t, x(t), r(t)) \in \Psi$, i.e.

$$u(t, r(t)) \triangleq \psi(t, x(t), r(t)) = K(r(t))x(t), \psi : [t_0, \infty) \times \mathbf{R}^n \times \mathbf{S} \to \mathbf{R}^m, \tag{3}$$

which minimizes the following infinite horizon performance criterion

$$V(t, x(t), r(t) = i, u_i) = \mathrm{E}\left\{\int_{t_0}^{\infty} [x^{\mathrm{T}}(t)Q(r(t))x(t) + u^{\mathrm{T}}(t)R(r(t))u(t)]\mathrm{d}t | x(t_0), r(t_0), \forall t_0 \geq 0\right\} \tag{4}$$

where $K(r(t))$ is a state feedback gain matrix to be determined, and the mode-dependent weighting matrices $Q(r(t))$ and $R(r(t))$ are real valued with $Q^{\mathrm{T}}(r(t)) = Q(r(t)) \geq 0$, $R^{\mathrm{T}}(r(t)) = R(r(t)) > 0$ $(\forall r(t) \in \mathbf{S})$. For convenience, we denote $A_0(r(t))$, $B(r(t))$, $A_k(r(t))$, $Q(r(t))$, $R(r(t))$ as A_{0i}, B_i, A_{ki}, Q_i, R_i, respectively, with $r(t) = i, i \in \mathbf{S}$.

Note that the mode-dependent control policy $u_i(t)$ not only stochastic stabilizes the Markovian jump Itô stochastic system on $\mathbf{R}^m$ but also guarantees that (4) is finite for all $i \in \mathbf{S}$. That is to say, mode-dependent control policy $u_i(t)$ must be admissible for all $i \in \mathbf{S}$. To this end, the following definition of *mode-dependent admissible control policy* is introduced.

Definition 1. *(Mode-Dependent Admissible Control Policy, [35]) For all $i \in \mathbf{S}$, the mode-dependent control policy $u_i(t) = \psi(t, x(t), i)$ with $\psi(t_0, x(t_0), r(t_0)) \equiv 0$ is said to be mode-dependent admissible with respect to (4) on Ψ, if $\psi(t, x(t), i)$ is continuous on a compact set $\Psi \subseteq \mathbf{R}^m$, stochastic stabilizes the continuous-time Markovian jump Itô stochastic system (1)-(2), and yields a **finite** mode-dependent value function $V_i(x(t), u_i(t)) = \mathrm{E}\{V(t, x(t), r(t) = i, u_i(t))\}|t_0, x(t_0), r(t_0)\}|_{\psi(t,x(t),i)}$.*

Then, the infinite horizon optimal control problem for continuous-time Markovian jump Itô stochastic system (1)-(2) is equivalent to find the following a set of mode-dependent admissible control policy:

$$u_i(t) = u_i^*(x(t)) \triangleq \arg\min_{u_i \in \Psi} V_i(x(t), u_i(t)), \forall i \in \mathbf{S}. \tag{5}$$

Definition 2. *The continuous-time Markovian jump Itô stochastic system (1)-(2) is asymptotically mean square stable (AMSS) if for any $x_0 \in \mathbf{R}^n$ and $r(0) \in \mathbf{S}$, there holds*

$$\lim_{t \to \infty} \mathrm{E}\{\|x(t)\|^2\} = 0,$$

where $x(t) = x(t, x_0, r(0))$ is a sample solution of the system.

Lemma 1. *(Lemma 2.4, [15]) With the linear control law (3), the continuous-time Markovian jump Itô stochastic system (2) is AMSS and satisfies (5) if and only if the following stochastic coupled algebraic Riccati equation (CARE)*

$$\mathcal{L}(P_i) = A_{0i}^{\mathrm{T}}P_i + P_iA_{0i} + \sum_{k=1}^{r} A_{ki}^{\mathrm{T}}P_iA_{ki} + \sum_{j=1}^{N}\pi_{ij}P_j + Q_i - P_iB_iR_i^{-1}B_i^{\mathrm{T}}P_i = 0 \tag{6}$$

has a unique solution $P = (P_1, \cdots, P_i, \cdots, P_N)$ *with* $P_i \geq 0$, $i \in \mathbf{S}$. *In this case, the optimal controller gain is given by*

$$K_i = -R_i^{-1}B_i^{\mathrm{T}}P_i,\ i \in \mathbf{S}. \tag{7}$$

2.2 Implicit iteration algorithm for stochastic CARE

It is seen from Lemma 1 that the optimal control problem for continuous-time Markovian jump Itô stochastic system (1)-(2) is reduced to find a set of stochastically stabilizing solutions P_i of the stochastic CARE (6). In [14], the authors proposed an iteration algorithm (i.e. so-called *implicit iteration algorithm*) for to learn the solution of continuous-time stochastic coupled Lyapunov equation of the continuous-time Markovian jump Itô stochastic system (2) with $u(t) = 0$. At first, we review the iteration algorithm.

The set of coupled Lyapunov equations corresponding to the system (2) with $u(t) = 0$ is given by

$$A_{0i}^{\mathrm{T}}P_i + P_iA_{0i} + \sum_{k=1}^{r} A_{ki}^{\mathrm{T}}P_iA_{ki} + \sum_{j=1}^{N}\pi_{ij}P_j + Q_i = 0. \tag{8}$$

The implicit iteration algorithm for (8) is given as follows [14]:

Algorithm 1 Implicit iteration algorithm for stochastic coupled Lyapunov equation

- *Step 1.* Given an initial condition $P(0) = (P_1(0), P_2(0), \cdots, P_N(0))$ and $\beta_i \geq 0$, $i \in \mathbf{S}$. Let $k = 0$.
- *Step 2.* Solve the following N standard continuous-time Lyapunov equations for $P_i(k+1)$:

$$\mathcal{A}_i^{\mathrm{T}}P_i(k+1) + P_i(k+1)\mathcal{A}_i = -\sum_{s=1}^{r} A_{si}^{\mathrm{T}}P_i(k)A_{si} - \sum_{j=1,j\neq i}^{N}\pi_{ij}P_j(k) - \beta_iP_i(k) - Q_i,\ i \in \mathbf{S},$$

  where $\mathcal{A}_i = A_{0i} + \frac{\pi_{ii}}{2}I - \frac{\beta_i}{2}I$.
- *Step 3.* Set $i = i+1$. If $\max_{i\in\mathbf{S}}(\|P_i(k) - P_i(k-1)\|) \leq \epsilon$ ($\epsilon > 0$ is a given small constant), stop and output $P_i(k)$ as the solution P_i of the stochastic Lyapunov equation (8), else, go to Step 2 and continue.

Next, we extend the implicit iteration algorithm to solve the stochastic CARE (6). The algorithm is presented as follows.

It is worth mentioning that the convergence proof of Algorithm 2 can not be obtained from Theorem 5 of [14] directly and required the future investigations. Nevertheless, when choosing $\beta_i = 0$ in Algorithm 2, the iterative equation (9) reduces to

$$\begin{aligned}&\left[A_{0i} + \frac{\pi_{ii}}{2}I - B_iR_i^{-1}B_i^{\mathrm{T}}P_i(k)\right]^{\mathrm{T}}P_i(k+1) + P_i(k+1)\left[A_{0i} + \frac{\pi_{ii}}{2}I - B_iR_i^{-1}B_i^{\mathrm{T}}P_i(k)\right]\\ &= -P_i(k)B_iR_i^{-1}B_i^{\mathrm{T}}P_i(k) - \sum_{s=1}^{r} A_{si}^{\mathrm{T}}P_i(k)A_{si} - \sum_{j=1,j\neq i}^{N}\pi_{ij}P_j(k) - Q_i,\ i \in \mathbf{S}.\end{aligned} \tag{10}$$

Algorithm 2 Implicit iteration algorithm for stochastic CARE

- *Step 1.* Given an initial stabilizing sequence $\{P(0)\}$ with $P(0) = (P_1(0), P_2(0), \cdots, P_N(0))$ and $\beta_i \geq 0$, $i \in \mathbf{S}$. Let $k = 0$.
- *Step 2.* Solve the following N standard continuous-time Lyapunov equations for $P_i(k+1)$:

$$\begin{aligned}&\left[\mathcal{A}_i - B_i R_i^{-1} B_i^{\mathrm{T}} P_i(k)\right]^{\mathrm{T}} P_i(k+1) + P_i(k+1)\left[\mathcal{A}_i - B_i R_i^{-1} B_i^{\mathrm{T}} P_i(k)\right]\\ &= -P_i(k) B_i R_i^{-1} B_i^{\mathrm{T}} P_i(k) - \sum_{s=1}^{r} A_{si}^{\mathrm{T}} P_i(k) A_{si} - \sum_{j=1, j\neq i}^{N} \pi_{ij} P_j(k) - \beta_i P_i(k) - Q_i,\ i \in \mathbf{S},\end{aligned} \tag{9}$$

  where $\mathcal{A}_i = A_{0i} + \frac{\pi_{ii}}{2} I - \frac{\beta_i}{2} I$.
- *Step 3.* Set $i = i + 1$. If $\max_{i\in\mathbf{S}}(\|P_i(k) - P_i(k-1)\|) \leq \epsilon$ ($\epsilon > 0$ is a given small constant), stop and output $P_i(k)$ as the solution P_i of the stochastic CARE (6), else, go to Step 2 and continue.

It is not difficult to find that the iterative algorithm (10) is similar to the so-called Lyapunov iteration algorithm in [40] that converged to the solutions of the CARE of the optimal control problem for continuous-time Markov jump linear systems. To this end, we can conclude the following convergence theorem for Algorithm 2 in this special case.

Assumption 1. *([15, 40]) The continuous-time Markovian jump Itô stochastic system (1)-(2) is stochastically stabilizable by the optimal feedback controls (5), and the pairs $(A_0, (A_1, \cdots, A_r), \sqrt{Q_i})$ is stochastically detectable.*

Theorem 1. *Suppose the Assumption 1 holds. Set $\beta_i = 0$ for all $i \in \mathbf{S}$. The Algorithm 2 converges to the unique solution $P^* = (P_1^*, P_2^*, \cdots, P_N^*)$ of the stochastic CARE (6) for any initial stabilizing sequence $P(0) = (P_1(0), P_2(0), \cdots, P_N(0))$.*

Proof. The proof is omitted here because it is a trivial extension of the convergence proof in [40]. □

3 ST-Based Data-Driven Policy Iteration Algorithm

In this section, we propose a new policy iteration algorithm that will solve online for the optimal control problem of the continuous-time Markovian jump Itô stochastic systems. The algorithm is based on an decoupling technique, i.e. subsystems transformation (ST) [35], and a two-step iteration framework.

3.1 ST: An new decoupling technique

Before introducing the ST theorem, we review the well-known Kleinman's algorithm for solving the continuous-time ARE at first. Consider the continuous-time linear system described by

$$\dot{x}(t) = Ax(t) + Bu(t). \tag{11}$$

It is well known that the solution to the infinite horizon optimal control problem of system (11) can be found by solving the following well-known algebraic Riccati equation (ARE)

$$A^{\mathrm{T}} P + PA + Q - PBR^{-1}B^{\mathrm{T}} P = 0 \tag{12}$$

with the optimal controller $u(t) = -K^* x(t) = -R^{-1}B^{\mathrm{T}} P^* x(t)$.

Lemma 2. *[38] Let K_0 be any stabilizing feedback gain matrix, and let $P(k)$ be the symmetric positive definite solution of the Lyapunov equation*

$$[A - BK(k)]^{\mathrm{T}} P(k) + P(k)[A - BK(k)] + Q + K^{\mathrm{T}}(k)RK(k) = 0 \tag{13}$$

where $K(k)$, with $k = 1, 2, \cdots$, are defined recursively by:

$$K(k) = R^{-1}B^{\mathrm{T}}P(k-1). \tag{14}$$

Then, the following properties hold:

1. *$A - BK(k)$ is Hurwitz,*
2. *$\lim_{k\to\infty} P(k) = P^*, \lim_{k\to\infty} K(k) = K^*$.*

It is seen from Lemma 2 that, by iteratively solving the Lyapunov equation (13) and updating control gain $K(k)$ by (14), the solution the the ARE (12) is numerically approximated. Inspired by the work of [41] and Lemma 2, we give the following theorem on subsystems transformation from system (2).

Theorem 2. *(ST)*

Construct the following N continuous-time linear subsystems from stochastic system (2):

$$\dot{x}_{i,t} = \left(A_{0i} + \frac{\pi_{ii}}{2}I\right)x_{i,t} + B_i u_{i,t},\ (i = 1, 2, \cdots, N), \tag{15}$$

where $x_{i,t}$ is the ith subsystem state, and $u_{i,t}$ is the ith subsystem input. Let $\{K_i(0)\}$ be the sequence of any stabilizing feedback gain matrices of the system (15), and let $\{P_i(k)\}$ be the kth step solutions of the Kleinman's iteration algorithm for the system (15) ***with each step the parameter $\mathbf{Q}_i$ is updated recursively by***

$$\mathbf{Q}_i(k) = Q_i + \sum_{j=1, j\neq i}^{N} \pi_{ij}P_j(k-1) + \sum_{s=1}^{r} A_{si}^{\mathrm{T}}P_i(k-1)A_{si}. \tag{16}$$

Then, the following properties hold for any $i \in \mathbf{S}$:

1. *$A_{0i} + \frac{\pi_{ii}}{2}I - B_iK_i(k)$ is Hurwitz.*
2. *$\lim_{k\to\infty} P_i(k) = P_i^\star$.*

where the $\{P_i^\star\}$ are the solutions of the stochastic CARE (6).

Proof. By Lemma 2, the kth iteration Lyapunov equation of system (15) can be rewritten as

$$\left[A_{0i} + \frac{\pi_{ii}}{2}I - B_iK_i(k)\right]^{\mathrm{T}} P_i(k) + P_i(k)\left[A_{0i} + \frac{\pi_{ii}}{2}I - B_iK_i(k)\right] + \mathbf{Q}_i(k) + K_i^{\mathrm{T}}(k)R_iK_i(k) = 0. \tag{17}$$

Substituting (14) and (16) into (17), it has

$$\begin{aligned}&\left[A_{0i} + \frac{\pi_{ii}}{2}I - B_iR_i^{-1}B_i^{\mathrm{T}}P_i(k-1)\right]^{\mathrm{T}} P_i(k) + P_i(k)\left[A_{0i} + \frac{\pi_{ii}}{2}I - B_iR_i^{-1}B_i^{\mathrm{T}}P_i(k-1)\right]\\ &= -P_i(k-1)B_iR_i^{-1}B_i^{\mathrm{T}}P_i(k-1) - \sum_{s=1}^{r} A_{si}^{\mathrm{T}}P_i(k-1)A_{si} - \sum_{j=1, j\neq i}^{N} \pi_{ij}P_j(k-1) - Q_i,\ i \in \mathbf{S}.\end{aligned} \tag{18}$$

It is easy to find that (18) just is Algorithm 2 when $\beta_i = 0$. Recalling Theorem 1 and Lemma 2, the proof is completed.

□

Remark 1. *The decoupling transformation methods proposed in [41] can exactly decompose weakly coupled linear systems into N independent subsystems. As a result, the set of nonlinear coupled algebraic equations can be solved by operating on some linear decoupled Sylvester's equations [42]. Inspired by these ideas and Kleinman's algorithm [38], an new decoupling approach has been introduced in [35] for optimal control of continuous-time Markov jump systems. It is found from Theorem 2 that, the solutions of stochastic CARE (6) can be attained by parallel solving N Kleinman's iteration equations for the decomposed linear systems (15). Actually, the coupling relation between N Markovian jump modes has been explicitly shown in $\frac{\pi_{ii}}{2}$ of the decoupled subsystem (15) and the equation (16) for $\mathbf{Q}_i$-updating.*

Remark 2. *It is noted that the decoupled subsystem (15) does not contain the Itô stochastic term, which means that ST technique also helps us to overcome the stochastic characteristic of the concerned system and thus the integral reinforcement learning approach (such as [32, 37]) can be utilized (see Theorem 3). The information of both Markovian jump process and Itô stochastic term have been taken into account by the parameter iteration equation (16), which shows the difference and contribution of this work.*

3.2 Policy iteration algorithm

Notice that the continuous-time Markovian jump Itô stochastic system (1)-(2) are decomposed into N interrelated continuous-time linear time-invariant subsystems in form of (15) by ST approach in Theorem 2. In the following, we will show how implicit iteration algorithm in Algorithm 2 can be made using online sensory data without the need to identify interval dynamics of the N subsystems.

The ST-based data-driven policy iteration algorithm can thus be summarized as follows.

It observes that the system matrices $\{A_i, i \in \mathbf{S}\}$ are not included in implementing this algorithm. But in the policy improvement equations (20), the exact knowledge of the system matrices $\{B_i\}$ and the transfer matrix elements $\{\pi_{ij}\}$ are still required for the iterations. Therefore, this proposed data-driven policy iteration algorithm is a partially mode-free adaptive dynamic programming algorithm.

The convergence of Algorithm 3 is guaranteed by the following theorem.

Theorem 3. *Assuming the system matrices $\{A_i\}$ are stablilizable. The ST-based data-driven policy iteration in Algorithm 3 is equivalent to the computation iteration in Algorithm 2 when $\beta_i = 0$.*

Proof. For the ith linear subsystem in form of (15), its closed-loop form at each iteration step k can be described by

$$\dot{x}_{i,t} = \left[A_{0i} + \frac{\pi_{ii}}{2} I + B_i K_i(k)\right] x_{i,t}. \tag{21}$$

Then, it has

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t} x_{i,t}^{\mathrm{T}} P_i(k) x_{i,t} \\ &= x_{i,t}^{\mathrm{T}} \left[\left(A_{0i} + \frac{\pi_{ii}}{2} I + B_i K_i(k)\right)^{\mathrm{T}} P_i(k) + P_i(k) \left(A_{0i} + \frac{\pi_{ii}}{2} I + B_i K_i(k)\right)\right] x_{i,t}. \end{aligned} \tag{22}$$

Integrating (22) from t to $t+T$ yields

$$\begin{aligned} &x_{i,t+T}^{\mathrm{T}} P_i(k) x_{i,t+T} - x_{i,t}^{\mathrm{T}} P_i(k) x_{i,t} \\ &= \int_t^{t+T} x_{i,\tau}^{\mathrm{T}} \left[\left(A_{0i} + \frac{\pi_{ii}}{2} I + B_i K_i(k)\right)^{\mathrm{T}} P_i(k) + P_i(k) \left(A_{0i} + \frac{\pi_{ii}}{2} I + B_i K_i(k)\right)\right] x_{i,\tau} \mathrm{d}\tau. \end{aligned} \tag{23}$$

Algorithm 3 Data-driven policy iteration algorithm for stochastic CARE

- *Step 1.* Given an initial stabilizing sequence $\{P(0)\}$ with $P(0) = (P_1(0), P_2(0), \cdots, P_N(0))$. Let $k = 0$.
- *Step 2.* For the reconstructed N subsystems in form of (15), simultaneous solving the following N policy iteration equations for $\{P_i(k), i \in \mathbf{S}\}$:

1. **Policy evaluation:**

$$\begin{cases} x_{1,t}^{\mathrm{T}} P_1(k) x_{1,t} = \int_t^{t+T} x_{1,\tau}^{\mathrm{T}} \left[\mathbf{Q}_1(k) + K_1^{\mathrm{T}}(k) R_1 K_1(k)\right] x_{1,\tau} \mathrm{d}\tau \\ \qquad\qquad + x_{1,t+T}^{\mathrm{T}} P_1(k) x_{1,t+T}, \\ \qquad\qquad \vdots \\ x_{i,t}^{\mathrm{T}} P_i(k) x_{i,t} = \int_t^{t+T} x_{i,\tau}^{\mathrm{T}} \left[\mathbf{Q}_i(k) + K_i^{\mathrm{T}}(k) R_i K_i(k)\right] x_{i,\tau} \mathrm{d}\tau \\ \qquad\qquad + x_{i,t+T}^{\mathrm{T}} P_i(k) x_{i,t+T}, \\ \qquad\qquad \vdots \\ x_{N,t}^{\mathrm{T}} P_N(k) x_{N,t} = \int_t^{t+T} x_{N,\tau}^{\mathrm{T}} \left[\mathbf{Q}_N(k) + K_N^{\mathrm{T}}(k) R_N K_N(k)\right] x_{N,\tau} \mathrm{d}\tau \\ \qquad\qquad + x_{N,t+T}^{\mathrm{T}} P_N(k) x_{N,t+T}, \end{cases} \tag{19}$$

2. **Policy improvement ($\forall i \in \mathbf{S}$):**

$$\begin{cases} \mathbf{Q}_i(k+1) = Q_i + \sum_{j=1, j\neq i}^{N} \pi_{ij} P_j(k) + \sum_{s=1}^{r} A_{si}^{\mathrm{T}} P_i(k) A_{si}, \\ K_i(k+1) = R_i^{-1} B_i^{\mathrm{T}} P_i(k). \end{cases} \tag{20}$$

- *Step 3.* Set $i = i + 1$. If $\max_{i\in\mathbf{S}}(\|P_i(k) - P_i(k-1)\|) \leq \varepsilon$ ($\varepsilon > 0$ is a given small constant), stop and output $P_i(k)$ as the solution P_i of the stochastic CARE (6), else, go to Step 2 and continue.

On the other hand, it follows from (19) that

$$\begin{aligned} & x_{i,t+T}^{\mathrm{T}} P_i(k) x_{i,t+T} - x_{i,t}^{\mathrm{T}} P_i(k) x_{i,t} \\ & = -\int_t^{t+T} x_{i,\tau}^{\mathrm{T}} \left[\mathbf{Q}_i(k) + K_i^{\mathrm{T}}(k) R_i K_i(k)\right] x_{i,\tau} \mathrm{d}\tau. \end{aligned} \tag{24}$$

Combining (23) and (24), it get

$$\left[A_{0i} + \frac{\pi_{ii}}{2} I + B_i K_i(k)\right]^{\mathrm{T}} P_i(k) + P_i(k) \left[A_{0i} + \frac{\pi_{ii}}{2} I + B_i K_i(k)\right] = -\left[\mathbf{Q}_i(k) + K_i^{\mathrm{T}}(k) R_i K_i(k)\right]. \tag{25}$$

By invoking the policy improvement (20), we obtain the iteration equation in Algorithm 2 ($\beta_i = 0$). This completes the proof.

□

According to Theorem 1 and Theorem 3, it finds readily that Algorithm 3 converges to the solutions of the stochastic CARE (6).

3.3 Online implementation in a parallel framework

In this subsection, we derive the online implement approach for the ST-based data-driven policy iteration algorithm in Algorithm 3. It is necessary to point that, at each iteration step, the states of N parallel subsystems

are required to be measured online simultaneously. To this end, we rewrite the term $x_{i,t}^{\mathrm{T}}P_i(k)x_{i,t}$ in (19) as

$$x_{i,t}^{\mathrm{T}}P_i(k)x_{i,t} = [\hat{p}_i(k)]^{\mathrm{T}}\hat{x}_i(t), \tag{26}$$

where $\hat{x}_i(t)$ denotes the Kronecker product quadratic polynomial basis vector with the elements $\{x_{i,h}(t)x_{i,l}(t)\}$ $(h = 1, 2, \cdots, n; l = h, h+1, \cdots, n)$, that is,

$$\hat{x}_i = [x_{i,1}^2, x_{i,1}x_{i,2}, \cdots, x_{i,1}x_{i,n}, x_{i,2}^2, x_{i,2}x_{i,3}, \cdots, x_{i,n}^2]; \tag{27}$$

and, $\hat{p}_i(k)$ denotes the a column vector by stacking the elements of the diagonal and upper triangular part of the symmetric matrix P_i into a vector where the off-diagonal elements are taken as $2p_{i,hl}$, that is,

$$\hat{p}_i = [p_{i,11}, 2p_{i,12}, \cdots, 2p_{i,1n}, p_{i,22}, 2p_{i,23}, \cdots, p_{i,nn}]^{\mathrm{T}}. \tag{28}$$

By (27) and (28), the ith policy evaluation in (19) can be written as

$$[\hat{p}_i(k)]^{\mathrm{T}}[\hat{x}_{i,t} - \hat{x}_{i,t+T}] = \int_t^{t+T} x_{i,\tau}^{\mathrm{T}}[\mathbf{Q}_i(k) + K_i^{\mathrm{T}}(k)R_iK_i(k)]x_{i,\tau}\mathrm{d}\tau, \tag{29}$$

Notice that the symmetric matrix P_i has $n(n+1)/2$ unknown independent elements. Hence, we need at least M $(\geq n(n+1)/2)$ independent equations to get $P_i(k)$ in (29). That is to say, for any ith subsystem in form of (15), one should sample M state vector in each time interval T. Then, the sequence $\{P_i(k)\}$ can be obtained by solving the following parallel least-square equations:

$$\begin{cases} \hat{p}_1(k) = [X_1X_1^{\mathrm{T}}]^{-1}X_1Y_1, \\ \qquad\vdots \\ \hat{p}_i(k) = [X_iX_i^{\mathrm{T}}]^{-1}X_iY_i, \\ \qquad\vdots \\ \hat{p}_N(k) = [X_NX_N^{\mathrm{T}}]^{-1}X_NY_N, \end{cases} \tag{30}$$

where

$$X_i = \left[\begin{array}{cccc} \hat{x}_i(t) - \hat{x}_i(t+\Delta t) & \hat{x}_i(t+\Delta t) - \hat{x}_i(t+2\Delta t) & \cdots & \hat{x}_i(t+(M-1)\Delta t) - \hat{x}_i(t+M\Delta t) \end{array}\right],$$
$$Y_i = \left[y_i^{(0)}(k)\ y_i^{(1)}(k)\ \cdots\ y_i^{(M-1)}(k)\right]^{\mathrm{T}},$$

with $\Delta t = T/M$ and $y_i^{(r)}(k) = \int_{t+r\Delta t}^{t+(r+1)\Delta t} x_{i,\tau}^{\mathrm{T}}\left[\mathbf{Q}_i(k) + K_i^{\mathrm{T}}(k)R_iK_i(k)\right]x_{i,\tau}\mathrm{d}\tau$ $(r = 0, 1, \cdots, M-1)$.

Remark 3. *In order to keep persistent excitation (PE) property of the data necessary to obtain the solution given by (30), one has to continue sampling the states information from overall N subsystems at each iterative step until the solution of the parallel least-squares sense is feasible. An effective method to prevent the PE missing is to add some small excitation signal (or some additive exploration noise) in the control input $u_i(t)$ [24, 32, 43]. Moreover, it is worth to point out that, before learning the ultimate optimal controller, the controller solutions obtained at each step are all suboptimal state-feedback controllers for the concerned continuous-time Itô stochastic systems with Markovian jumps.*

Remark 4. *It is worth to point out that when taking the dynamic uncertainties or external disturbances into consideration, the robust adaptive dynamic programming approach in [33] or ADP-based H_∞ control strategy in [44] may be employed to solve the corresponding adaptive optimal control problem for the continuous-time* ***uncertain*** *stochastic Markovian jumping systems.*

4 Simulation Studies

In this section, we borrow a fourth-order continuous-time stochastic system with Markovian jumps (A modified version of Example 2 in [14]) to illustrate the feasibility and validity of the proposed iteration algorithms.

Consider the stochastic system in form of (1)-(2) with $N = 2$, $r = 1$, and

$$\Pi = \begin{bmatrix} -0.6 & 0.6 \\ 1 & -1 \end{bmatrix},$$

$$A_{01} = \begin{bmatrix} -1.000 & -1.0000 & -2.0000 & 1.0000 \\ -0.6667 & -3.5000 & 1.0000 & 1.1670 \\ 1.0000 & 0.5000 & -3.0000 & 0.5000 \\ -2.0000 & -2.5000 & 1.0000 & -2.5000 \end{bmatrix}, \quad A_{02} = \begin{bmatrix} -1.3333 & 1.0000 & -2.0000 & 0.3333 \\ 0.0000 & -3.5000 & 1.0000 & 1.5000 \\ 1.0000 & 1.5000 & -4.0000 & -0.5000 \\ -1.3333 & -3.5000 & 1.0000 & -1.1667 \end{bmatrix},$$

$$A_{11} = \begin{bmatrix} 0.9003 & 0.7826 & 0.6428 & 0.8436 \\ -0.5377 & 0.5242 & -0.1106 & 0.4764 \\ 0.2137 & -0.0871 & 0.2309 & -0.6475 \\ -0.0280 & -0.9630 & 0.5839 & -0.1886 \end{bmatrix}, \quad A_{12} = \begin{bmatrix} 0.8709 & -0.8842 & -0.7222 & -0.4556 \\ 0.8338 & -0.2943 & -0.5945 & -0.6024 \\ -0.1795 & 0.6263 & -0.6026 & -0.9695 \\ 0.7873 & -0.9803 & 0.2076 & 0.4936 \end{bmatrix},$$

$$B_1 = B_2 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \ Q_1 = Q_2 = I_4, \ R_1 = R_2 = 1.$$

4.1 Algorithm 2: Implicit iteration algorithm for stochastic CARE

Let the value of stop criterion ϵ in Algorithm 2 as 10^{-10} and $\beta_1 = \beta_2 = 0$. With the initial conditions $P_1(0) = P_2(0) = 0$, the following solutions of the stochastic CARE (6) have been obtained after 27 iterations

$$P_1^\star = P_1(27) = \begin{bmatrix} 0.5779 & -0.1139 & -0.2107 & 0.0114 \\ -0.1139 & 0.3006 & 0.0910 & -0.0199 \\ -0.2107 & 0.0910 & 0.3964 & 0.0820 \\ 0.0114 & -0.0199 & 0.0820 & 0.3333 \end{bmatrix},$$

$$P_2^\star = P_2(27) = \begin{bmatrix} 0.5136 & -0.0140 & -0.2340 & -0.0760 \\ -0.0140 & 0.3999 & 0.0397 & -0.1538 \\ -0.2340 & 0.0397 & 0.3272 & 0.0962 \\ -0.0760 & -0.1538 & 0.0962 & 0.3561 \end{bmatrix}. \tag{31}$$

Moreover, it shows that

$$\|P_1(27) - P_1(26)\| = 6.2941 \times 10^{-11}, \ \|P_2(27) - P_2(26)\| = 6.2861 \times 10^{-11}. \tag{32}$$

The matrix parameters in P_1 and P_2 after each iteration are plotted in Figures 1 and 2, respectively. It can be seen that the convergence of Algorithm 2 is well.

4.2 Algorithm 3: Data-driven policy iteration algorithm for stochastic CARE

For the purpose of demonstrating the Algorithm 3, the initial states of the two subsystems are taken as $x_1(0) = x_2(0) = \begin{bmatrix} 0.1 & 0.1 & 0.1 & 0 \end{bmatrix}^{\mathrm{T}}$, and the initial parameters are selected as $P_1(0) = P_2(0) = 0.1 I_{4\times 4}$. In the simu-

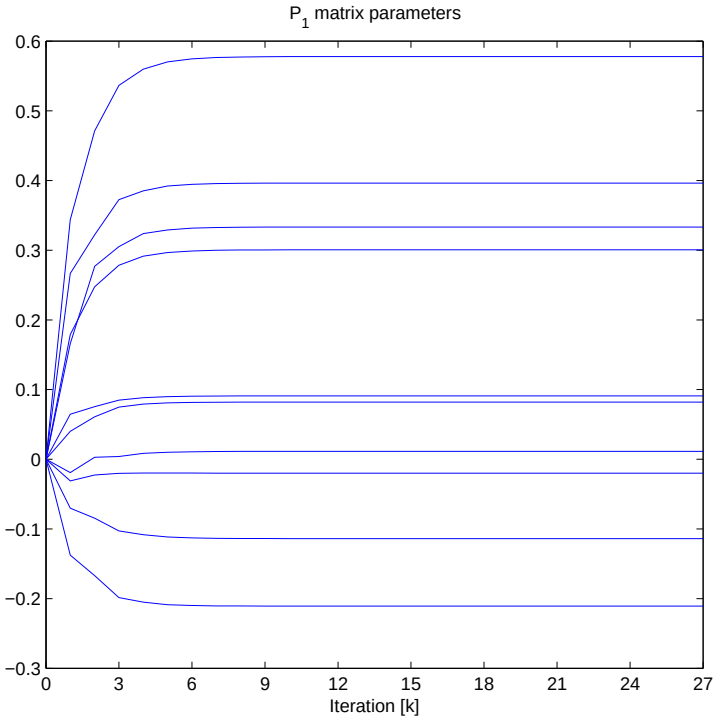

Figure 1: P_1 matrix parameters updated at each iteration step in Algorithm 2.

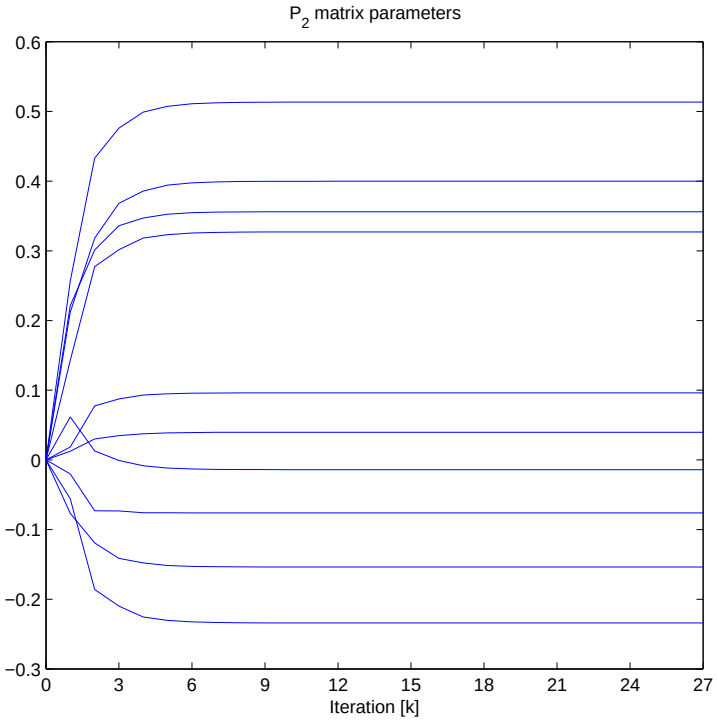

Figure 2: P_2 matrix parameters updated at each iteration step in Algorithm 2.

lations, the optimal controllers are updated every $0.6s$ and the two subsystems states information are conducted over each interval of $0.06s$. One thus have 10 sample data for every subsystem in each update interval. It should be noted that there are 10 independent elements in the symmetric matrices $P_1(k)$ and $P_2(k)$. Hence, at each update interval $0.6s$ and for every reconstructed subsystem, the 10 sample data are enough for exciting to solve parallel least-square equations (35). In addition, the parameter in Algorithm 3 is given as $\varepsilon = 10^{-13}$.

After 5 iteration steps, we obtain the following solutions by Algorithm 3:

$$P_1(5) = \begin{bmatrix} 0.5707 & -0.1139 & -0.2097 & 0.0102 \\ -0.1139 & 0.2917 & 0.0903 & -0.0187 \\ -0.2097 & 0.0903 & 0.3967 & 0.0814 \\ 0.0102 & -0.0187 & 0.0814 & 0.3283 \end{bmatrix},$$

$$P_2(5) = \begin{bmatrix} 0.5087 & -0.0126 & -0.2312 & -0.0759 \\ -0.0126 & 0.3949 & 0.0387 & -0.1519 \\ -0.2312 & 0.0387 & 0.3270 & 0.0962 \\ -0.0759 & -0.1519 & 0.0962 & 0.3536 \end{bmatrix}. \quad (33)$$

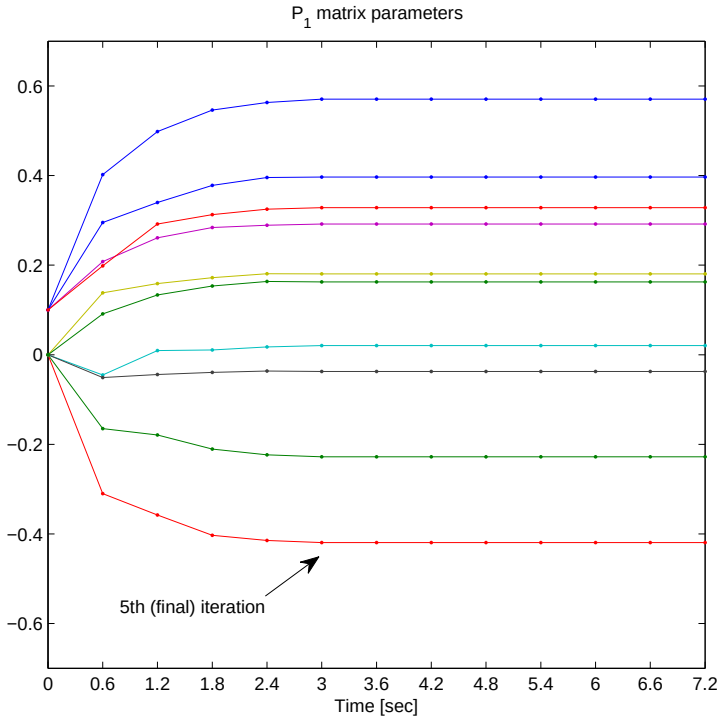

Figure 3: P_1 matrix parameters updated at each iteration step in Algorithm 3.

The simulation results are shown in Figures 3-6. Among them, Figures 3 and 4 show the matrix parameters in $P_1(k)$ and $P_2(k)$ after each iteration in Algorithm 3, respectively. It can be seen that, after five iterations, we obtain $\max\{\|P_1(5) - P_1^\star\|, \|P_2(5) - P_2^\star\|\} = 0.0093$, which illustrated that the proposed Algorithm 3 can converge to the solutions of the stochastic CARE (6) with a satisfied accuracy. Figures 5 and 6 depict the state evolutions during the simulation for the two subsystems in (15), respectively.

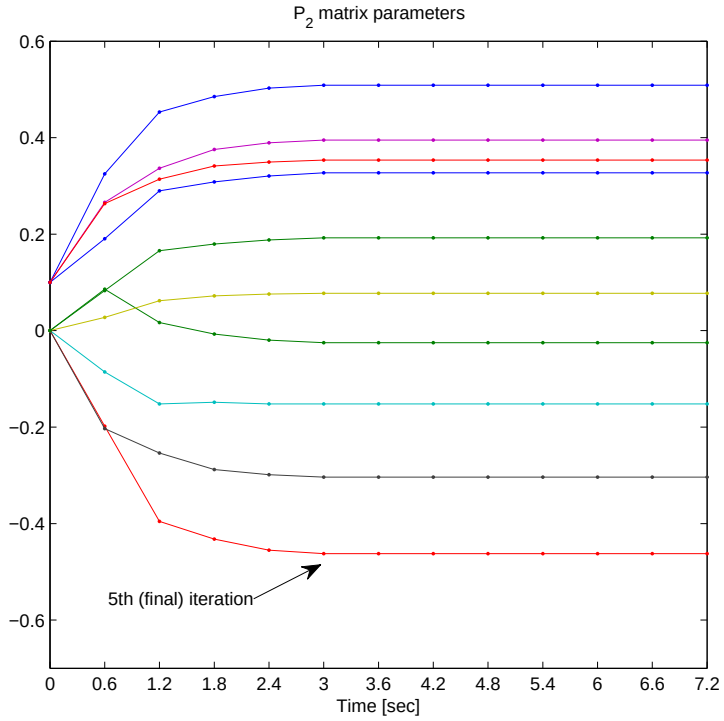

Figure 4: P_2 matrix parameters updated at each iteration step in Algorithm 3.

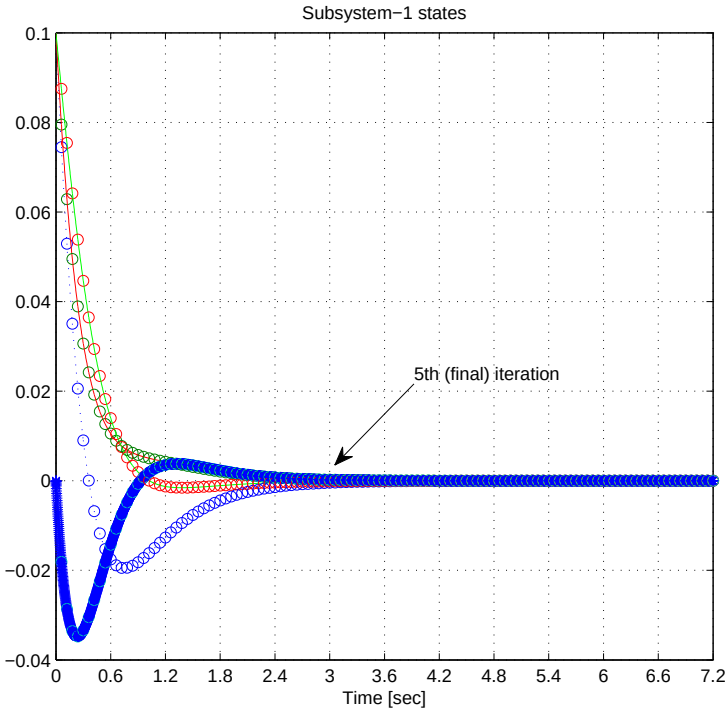

Figure 5: State responses of Subsystem-1 during the simulation.

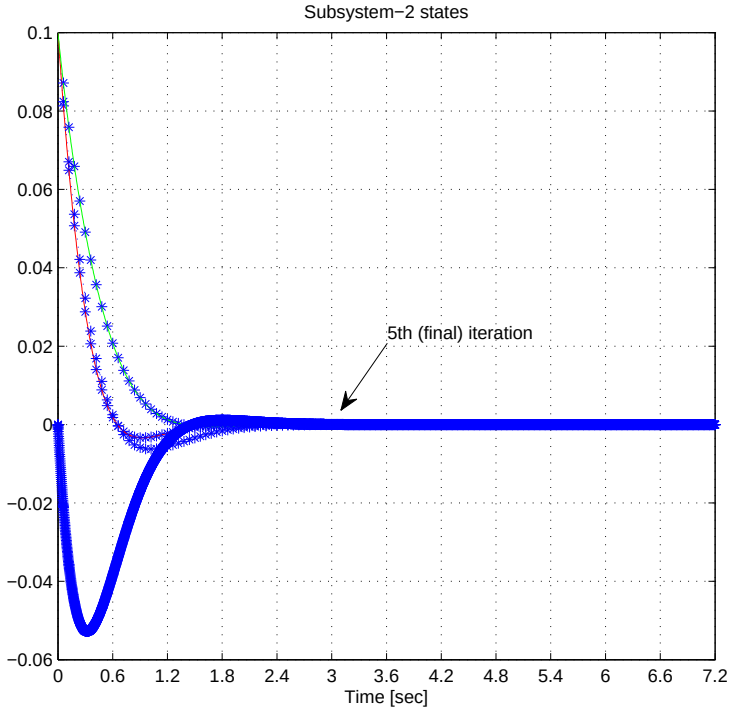

Figure 6: State responses of Subsystem-2 during the simulation.

5 Conclusions

This work proposed a data-driven policy iteration algorithm to solve the infinite horizon optimal control problem for the continuous-time Itô stochastic systems with Markovian jumps. By applying subsystems transformation technique, the optimal solutions of the stochastic coupled algebraic Riccati equation can be learned effectively from the decomposed subsystems data. It has been proved that the convergence of the proposed policy iteration algorithm is equivalent to the implicit iteration algorithm presented in [14]. It is interesting for us to extend the proposed approaches to time-delay stochastic systems with Markovian jumps [45, 46, 47, 48, 49] in a near future.

References

[1] Niu, Y., Ho, D.W.C., Wang, X.: 'Sliding mode control for Itô stochastic jump systems with Markovian switching', *Automatica*, 2007, **43**, (10), pp. 1784-1790

[2] Chen, B., Niu, Y., Zou, Y.: 'Adaptive sliding mode control for stochastic Markovian jumping systems with actuator degradation', *Automatica*, 2013, **49**, (6), pp. 1748-1754

[3] Chen, B., Niu, Y., Zou, Y.: 'Sliding mode control for stochastic Markovian jumping systems with incomplete transition rate', *IET Control Theory Appl.*, 2013, **7**, (10), pp. 1330-1338

[4] Chen, Y., Zheng, W.X.: 'Exponential H_∞ filtering for stochastoc Markovian jump systems with time delays', *Int. J. Robust Nonlinear Control*, 2014, **24**, (4), pp. 625-643

[5] Wang, Z., Liu, Y., Liu, X.: 'Exponential stabilization of a class of stochastic system with Markovian jump parameters and mode-dependent mixed time-delays', *IEEE Trans. Autom. Control*, 2010, **55**, (7), pp. 1656-1662

[6] Mao, X., Yuan, C.: '*Stochastic Differential Equations with Markovian Switching*'. London, U.K.: Imperial College Press, 2006

[7] Dragan, V., Morozan, T.: 'The linear quadratic optimization problems for a class of linear stochastic systems with multiplicative white noise and Markovian jumping', *IEEE Trans. Autom. Control*, 2004, **49**, (5), pp. 665-675

[8] Chen, Y., Zheng, W.X.: 'Stochastic state estimation for neural networks with distributed delays and Markovain jump', *Neural Netw.*, 2012, **25**, pp. 14-20

[9] Zhang, H., Guan, Z., Feng, G.: 'Reliable dissipative control for stochastic impulsive systems', *Automatica*, 2008, **44**, (4), pp. 1004-1010

[10] Zhang, L.: 'H_∞ estimation for discrete-time piecewise homogeneous Markov jump linear systems', *Automatica*, 2009, **45**, (11), pp. 2570-2576

[11] Yan, H., Su, Z., Zhang, H., Yang, F.: 'Observer-based H_∞ control for discrete-time stochastic systems with quantisation and random communication delays', *IET Control Theory Appl.*, 2013, **7**, (3), pp. 372-379

[12] Borno, I.: 'Parallel computation of the solutions of coupled algebraic Lyapunov equations', *Automatica*, 1995, **31**, (9), pp. 1345-1347

[13] Borno, I., Gajic, Z.: 'Parallel algorithm for solving coupled algebraic Lyapunov equations of discrete-time jump linear systems', *Comput. Math. Appl.*, 1995, **30**, (7), pp. 1-4

[14] Li, Z., Zhou, B., Lam, J., Wang, Y.: 'Positive operator based iterative algorithms for solving Lyapunov equations for Itô stochastic systems with Markovian jumps', *Appl. Math. Comput.*, 2011, **217**, (21), pp. 8179-8195

[15] Huang, Y., Zhang, W., Feng, G.: 'Infinite horizon H_2/H_∞ control for stochastic systems with Markovian jumps', *Automatica*, 2008, **44**, (3), pp. 857-863

[16] Sutton, R.S., Barto, A.G.: '*Reinforcement Learning: An Introduction*'. MIT Press, Cambridge MA, 1998

[17] Lewis, F.L., Vrabie, D.: 'Reinforcement learning and adaptive dynamic programming for feedback control', *IEEE Trans. Circuits Syst. Mag.*, 2009, **9**, (3), pp. 32-50

[18] Kiumarsi, B., Lewis, F.L., Levine, D.S.: 'Optimal control of nonlinear discrete time-varying systems using a new neural network approximation structure', *Neurocomputing*, 2015, **156**, pp. 157-165

[19] Vrabie, D., Lewis, F.: '*Online Adaptive Optimal Control Based On Reinforcement Learning*'. Optimization and Optimal Control, Springer New York, 2010, pp. 309-323

[20] Lewis, F.L., Liu, D.: '*Reinforcement Learning and Approximate Dynamic Programming for Feedback Control*'. Wiley, Hoboken, NJ, 2013

[21] Murray, J.J., Cox, C.J., Lendaris, G.G., Saeks, R.: 'Adaptive dynamic programming', *IEEE Trans. Syst. Man Cybern. C, Appl. Rev.*, 2002, **32**, (2), pp. 140-153

[22] Wang, F., Zhang, H., Liu, D.: 'Adaptive dynamic programming: An introduction', *IEEE Trans. Comput. Intell. Mag.*, 2009, **4**, (2), pp. 39-47

[23] He, P., Jagannathan, S.: 'Reinforcement learning neural-network-based controller for nonlinear discrete-time systems with input constraints', *IEEE Trans. Syst. Man Cyber. B, Cyber.*, 2007, **37**, (2), pp. 425-436

[24] Vamvoudakis, K.G., Lewis, F.: 'Online solution of nonlinear two-player zero-sum games using synchronous policy iteration', *Int. J. Robust Nonlinear Control*, 2012, **22**, (13), pp. 1460-1483

[25] Zhang, H., Lewis, F.: 'Adaptive cooperative tracking control of higher-order nonlinear systems with unknown dynamics', *Automatica*, 2012, **48**, (7), pp. 1432-1439

[26] Zhang, H., Wei, Q., Luo, Y.: 'A novel infinite-time optimal tracking control scheme for a class of discrete-time nonlinear systems via the greedy HDP iteration algorithm', *IEEE Trans. Syst. Man Cybern. B, Cybern.*, 2008, **38**, (4), pp. 937-942

[27] Bhasin, S., Kamalapurkar, R., Johnson, M., Vamvoudakis, K.G., Lewis, F., Dixon, W.E.: 'A novel actor-critic-identifier architecture for approximate optimal control of uncertain nonlinear systems', *Automatica*, 2013, **49**, (1), pp. 82-92

[28] Al-Tamimi, A., Lewis, F., Abu-Khalaf, M.: 'Discrete-time nonlinear HJB solution using approximate dynamic programming: Convergence proof', *IEEE Trans. Syst. Man Cybern. B, Cybern.*, 2008, **38**, (4), pp. 943-949

[29] Luo, B., Huang, T., Wu, H., Yang, X.: 'Data-driven H_∞ control for nonlinear distributed parameter systems', *IEEE Trans. Neural Netw. Learning Syst.*, 2015, **26**, (11), pp. 2949-2961

[30] Li, C., Liu, D., Li, H.: 'Finite horizon optimal tracking control of partially unknown linear continuous-time systems using policy iteration', *IET Control Theory Appl.*, 2015, **9**, (12), pp. 1791-1801

[31] Wu, H., Luo, B.: 'Simultaneous policy update algorithms for learning the solution of linear continuous-time H_∞ state feedback control', *Inf. Sci.*, 2013, **222**, pp. 472-485

[32] Jiang, Y., Jiang, Z.: 'Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics', *Automatica*, 2012, **48**, (10): pp. 2699-2704

[33] Jiang, Y., Jiang, Z.: 'A robust adaptive dynamic programming pronciple for sensorimotor control with signal-dependent noise', *J. Syst. Sci. Complex.*, 2014, **28**, (2), pp. 261-288

[34] Zhong, X., He, H., Zhang, H., Wang, Z.: 'Optimal control for unknown discrete-time nonlinear Markov jump systems using adaptive dynamic programming', *IEEE Trans. Neural Netw. Learning Syst.*, 2014, **25**, (12), pp. 2141-2155

[35] He, S., Song, J., Ding, Z., Liu, F.: 'Online adaptive optimal control for continuous-time Markov jump linear systems using a novel policy iteration algorithm', *IET Control Theory Appl.*, 2015, **9**, (10), pp. 1536-1543

[36] Jiang, Y., Jiang, Z.: 'Approximate dynamic programming for optimal stationary control with control-dependent noise', *IEEE Trans. Neural Netw.*, 2011, **22**, (12), pp. 2392-2398

[37] Vrabie, D., Pastrvanu, O., Abu-Khalaf, M., Lewis F.: 'Adaptive optimal control for continuous-time linear systems based on policy iteration', *Automatica*, 2009, **45**, (2), pp. 477-484

[38] Kleinman, D.: 'On an iterative technique for Riccati equation computations', *IEEE Trans. Autom. Control*, 1968, **13**, (1), pp. 114-115

[39] Ji, Y., Chizeck, H.J.: 'Controllability, stabilizability, and continuous-time Markovian jump linear quadratic control', *IEEE Trans. Autom. Control*, 1990, **35**, (7), pp. 777-788

[40] Gajic, Z., Borno, I.: 'Lyapunov iterations for optimal control of jump linear systems at steady state', *IEEE Trans. Autom. Control*, 1995, **40**, (11), pp. 1971-1975

[41] Gajic, Z., Borno, I.: 'General transformation for block diagonalization of weakly coupled linear systems composed of N-subsystems', *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, 2000, **47**, (6), pp. 909-912

[42] Gajic, Z., Shen, X.: 'Decoupling transformation for weakly coupled linear systems', *Int. J. Control*, 1989, **50**, (4), pp. 1517-1523

[43] Lee, J.Y., Park, J.B., Choi, Y.H.: 'Policy-iteration-based adaptive optimal control for uncertain continuous-time linear systems with excitation signals', *in Proc. IEEE Int. Conf. Control Autom. Syst. (ICCAS)*, 2010, pp. 646-651

[44] Song, J., He, S., Ding, Z., Liu, F.: 'A new iterative algorithm for solving H_∞ control problem of continuous-time Markovain jumping linear systems based on online implementation', *Int. J. Robust Nonlinear Control*, 2016, DOI: 10.1002/rnc.3531

[45] Zhang, Y.: 'Stability of discrete-time delay Markovian jump systems with stochastic non-linearity and impulses', *IET Control Theory Appl.*, 2013, **7**, (18), pp. 2178-2187

[46] Zhang, Y.: 'Stochastic stability of discrete-time Markovian jump delay neural networks with impulses and incomplete information on transition probability', *Neural Netw.*, 2013, **46**, pp. 276-282

[47] Zhang, Y.: 'Robust stochastic stability of uncertain discrete-time impulsive Markovain jump delay systems with multiplicative noises', *Int. J. Syst. Sci.*, 2015, **46**, (12), pp. 2210-2220

[48] Zhang, H., Yan, H., Liu, T., Chen, Q.: 'Fuzzy controller design for nonlinear impulsive fuzzy systems with time delay', *IEEE Trans. Fuzzy Syst.*, 2011, **19**, (5), pp. 844-856

[49] Zhang, H., Feng, G., Yan, H., Chen, Q.: 'Sampled-data control of nonlinear networked systems with time-delay and quantization', *Int. J. Robust Nonlinear Control*, 2016, **26**, (5), pp. 919-933