Available online at www.sciencedirect.com

ScienceDirect

Journal of the Franklin Institute 360 (2023) 2361–2378

www.elsevier.com/locate/jfranklin

Off-policy reinforcement learning for tracking control of discrete-time Markov jump linear systems with completely unknown dynamics

Zhen Huang[a], Yidong Tu[a], Haiyang Fang[b], Hai Wang[c], Liang Zhang[a], Kaibo Shi[d], Shuping He[a,*]

[a] *Key Laboratory of Intelligent Computing and Signal Processing (Ministry of Education), School of Electrical Engineering and Automation, Hefei 230601, China*

[b] *Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, Hong Kong*

[c] *Discipline of Engineering and Energy; Centre for Water, Energy & Waste, Harry Butler Institute, Murdoch University, Perth, WA6150, Australia*

[d] *School of Electronic Information and Electrical Engineering, Chengdu University, Chengdu 610106, China*

Received 16 July 2022; received in revised form 28 September 2022; accepted 28 October 2022
Available online 9 November 2022

Abstract

In this paper, a model-free off-policy reinforcement learning (RL) algorithm is proposed to address the optimal tracking control (OTC) problem for discrete-time Markov jump linear systems (MJLSs). The tracking reference signal is firstly augmented into discrete-time MJLSs, whereby the original tracking control problem is converted to the optimal control problem of the augmented system. The corresponding augmented coupled game algebraic Riccati equation (ACGARE) is then derived. On this basis, an online RL algorithm is developed to solve the OTC problem by using the policy iteration (PI) technique. Then, a novel model-free method is proposed, which eliminates the requirement of the system dynamics and transition probability. Finally, a simulation example is provided to prove the convergence and validate the effectiveness of the proposed algorithm.

© 2022 The Franklin Institute. Published by Elsevier Ltd. All rights reserved.

* Corresponding author.
E-mail address: shuping.he@ahu.edu.cn (S. He).

https://doi.org/10.1016/j.jfranklin.2022.10.052
0016-0032/© 2022 The Franklin Institute. Published by Elsevier Ltd. All rights reserved.

1. Introduction

Over the last decades, Markov jump linear systems (MJLSs) have attracted tremendous attention, due to its potential applications in networked control [1], medical prognosis [2] and power circuit systems [3], etc. In the actual MJLSs, one key challenge in controlling is the existence of transition probabilities of the jumping process [4,5]. In general, it is difficult in practice to acquire knowledge of these probabilities and dynamics. Although some recent publications, such as synchronization [6,7], robust control [8,9] and finite-time control [10,11] have proposed different solutions for MJLSs with partially unknown transition probabilities, the relevant control problems with completely unknown probabilities and dynamics are still challenging. And in this paper, we will try to develop a fully model-free algorithm for MJLSs.

The optimal tracking, aiming at controlling the system states to track a predefined reference signal and minimizing the tracking error [12,13], has been more and more important in industrial scenarios, such as single-link robot arms [14], underwater vehicles [15] and sensor networks [16], etc. Adaptive dynamic programming (ADP) schemes [17–19] and reinforcement learning (RL) methods [20,21] have been effectively utilized in the optimal control designs. Traditionally, the solutions of the ADP-based tracking problems consist of two components: the steady-state control input and the feedback control input [22,23]. However, most of these traditional approaches require perfect information of system dynamics. An augmentation technique is therefore introduced for the optimal tracking control (OTC) problem such that the complete information of system dynamics is not necessarily provided. In recent years, by augmentation techniques, some ADP-based approaches have been studied to address the OTC problem [24,25] with partially or completely unknown dynamics.

On the other hand, within the research scope of iteration RL methods, policy iteration (PI) algorithm [26–28] is a main iteration way. Since the essence of the OTC problem is to solve the algebraic Riccati equation (ARE), the PI technique therefore starts with a non-optimal control policy and iteratively alternates between evaluation and improvement steps to find the optimal solution of the ARE. In recent decades, the PI technique has been widely applied to the OTC problem [29,30]. However, despite these achievements, very few learning algorithms have been proposed for the OTC problem of MJSs. To our best knowledge, the OTC problem for continuous-time MJLSs was studied in [31] without using the transition probability and partial dynamics. Another OTC problem solution for nonlinear discrete-time MJSs was presented without requiring the system dynamics [32], where the ADP-learning process was implicitly implemented by neural networks. However, the theoretical analysis of stability and convergence needs further investigation. In fact, for the OTC problem of MJSs, the elimination of the dependence on system dynamics and transition probability has not been solved yet. In this paper, we attempt to use the model-free off-policy RL algorithm for solving such OTC problem in MJSs.

To obviate the requirement for the system dynamics, an off-policy method based on RL was proposed [33,34]. The off-policy algorithm not only solves the optimal solution of complex equations but also avoids the inaccuracy that may be caused by the system identification. For solving the OTC problem, a model-free off-policy algorithm was presented [35].

Motivated by aforementioned discussions, we will develop a model-free off-policy RL algorithm to address the OTC problem for MJLSs without using the any information of system dynamics. In this paper, we firstly construct an augmented system composed of the original system dynamics and the reference trajectory dynamics. Then, an augmented coupled game

algebraic Riccati equation (ACGARE) is derived, i.e., the tracking problem is converted to the optimal control problem based on the augmentation technique. Finally, a model-free approach is presented to solve ACGARE without requiring the information of system dynamics and transition probability.

The main contributions can be listed as follows:

(1) By building an augmented system, the OTC problem of the original system is converted to the optimal control problem of the augmented system. This augmentation technique can reduce the complexity of calculation.
(2) A fully model-free off-policy RL algorithm is proposed to solve the ACGARE for MJLSs, without using any information of the system dynamics and transition probability.
(3) The stochastic stability and convergence of the developed algorithms for the OTC problem of discrete-time MJLSs are still proved. Besides, a simulation result is presented to verify the correctness and effectiveness of the proposed algorithm.

The structure of this paper can be arranged into five parts. In Section II, we introduce the discrete-time MJLSs, build the augmented system, and derive the ACGARE by the Hamiltonian-Bellman equation. In Section III, two RL methods are proposed to solve the relevant OTC problem. In Section IV, a simulation example is shown to verify the validity of the proposed algorithms. Finally, a brief conclusion is provided in Section V.

2. Problem formulation

2.1. System statement

Consider the following discrete-time MJLSs as

$$\begin{cases} x_{k+1} = A(\tau(k))x_k + B(\tau(k))u_k \\ y_k = C(\tau(k))x_k \end{cases} \tag{1}$$

where $x_k \in \Re^n$ is the state vector, $u_k \in \Re^m$ is the controlled input, $y_k \in \Re^q$ is the measured output. $A(\tau(k))$, $B(\tau(k))$ and $C(\tau(k))$ are coefficient matrices. $\tau(k)$ is a discrete-time Markov chain, which belongs to a finite set $S = \{1, 2, \cdots, m\}$. We assume that $\tau(k)$ is independent of the initial state x_0. The transition probability matrix is described as

$$P_{rob} = \begin{bmatrix} \pi_{11} & \pi_{12} & \cdots & \pi_{1m} \\ \pi_{21} & \pi_{22} & \cdots & \pi_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \pi_{m1} & \pi_{m2} & \cdots & \pi_{mm} \end{bmatrix}, \tag{2}$$

and the elements in (2) can be represented as

$$\pi_{ij} = P_{rob}\{\tau(k+1) = j | \tau(k) = i\}, \tag{3}$$

which indicates the jumping probability from the current mode i to the next mode j. Note that $\pi_{ij} \geq 0$, $\forall i, j \in S$ and $\sum_{j=1}^{m} \pi_{ij} = 1$. For simplification, $A(\tau(k))$, $B(\tau(k))$, $C(\tau(k))$, $Q(\tau(k))$ and $R(\tau(k))$ can be equivalently represented by A_i, B_i, C_i, Q_i and R_i for $\tau(k) = i$ in the following sections.

For discrete-time tracking problems, our control purpose is to generate the optimal strategy such that not only the system output can be ensured to track the reference signal, but also

the system stability is guaranteed while minimizing the following performance index as

$$J(x_k, r_k) = E\left[\sum_{z=k}^{\infty} U_z\right] = E\left[\sum_{z=k}^{\infty}[(C_i x_z - r_z)^{\mathrm{T}} Q_i (C_i x_z - r_z) + u_z^{\mathrm{T}} R_i u_z]\right], \tag{4}$$

where $Q_i > 0$ and $R_i \geq 0$ are two symmetric weighted matrices. U_z is the utility function at time step z.

2.2. Augmented system for the OTC problem

Before moving on, the following assumption is important to the main results.

Assumption 1. The reference signal r_k can be tracked by the system (1), which is described as

$$r_{k+1} = F r_k, \tag{5}$$

where F is the drift dynamic.

For system (1) and the desired trajectory (5), we can construct the augmented system as

$$X_{k+1} = \begin{bmatrix} x_{k+1} \\ r_{k+1} \end{bmatrix} = \begin{bmatrix} A_i & 0 \\ 0 & F \end{bmatrix} \begin{bmatrix} x_k \\ r_k \end{bmatrix} + \begin{bmatrix} B_i \\ 0 \end{bmatrix} u_k \triangleq T_i X_k + B_{1i} u_k, \tag{6}$$

where $X_k = \begin{bmatrix} x_k^T & r_k^T \end{bmatrix}^T$ is the aggregated state of the augmented system and

$$T_i = \begin{bmatrix} A_i & 0 \\ 0 & F \end{bmatrix}, B_{1i} = \begin{bmatrix} B_i \\ 0 \end{bmatrix}.$$

Considering the controlled input $u_k = -K_i X_k$, Eq. (4) can be rewritten as

$$\begin{aligned} J_k = V(X_k) &= E\left[\sum_{z=k}^{\infty}(X_z^{\mathrm{T}} Q_{1i} X_z + u_z^{\mathrm{T}} R_i u_z)\right] \\ &= E\left[\sum_{z=k}^{\infty}(X_z^{\mathrm{T}}(Q_{1i} + K_i^T R_i K_i) X_z\right] \\ &= E(X_k^{\mathrm{T}} P_i X_k), \end{aligned} \tag{7}$$

where $Q_{1i} = C_{1i}^{\mathrm{T}} Q_i C_{1i}$ with $C_{1i} = \begin{bmatrix} C_i & -I \end{bmatrix}$.

Remark 1. By constructing the augmented system (6), the tracking problem of the initial system is converted to the optimal control problem of the augmented system. This augmentation technique greatly reduces the complexity of calculation.

Remark 2. The existing approaches of the OTC problem usually require the system dynamics [22,29]. Therefore they cannot be directly applied to these systems whose dynamics are both prior unknown and unavailable during the controlling process. Especially for stochastic systems, the transition probabilities of these systems are difficult to obtain in practice.

Regarding those problems, we will develop an off-policy RL algorithm based on the augmentation technique, without using any prior knowledge of the system dynamics and transition probability.

2.3. Bellman equation and ACGARE for the OTC problem

The following theorem derives the Bellman equation and ACGARE of (7), which will be the basis of our main results.

Theorem 1. *After augmentation, the OTC problem for the augmented system (6) can be solved by using the following ACGARE*

$$\begin{aligned} P_i =& T_i^{\mathrm{T}}\left(\sum_{j=1}^{m}\pi_{ij}P_j\right)T_i + Q_{1i} - T_i^{\mathrm{T}}\left(\sum_{j=1}^{m}\pi_{ij}P_j\right)B_{1i}\Bigg[B_{1i}^{\mathrm{T}} \\ &\times\left(\sum_{j=1}^{m}\pi_{ij}P_j\right)B_{1i} + R_i\Bigg]^{-1}B_{1i}^{\mathrm{T}}\left(\sum_{j=1}^{m}\pi_{ij}P_j\right)T_i. \end{aligned} \tag{8}$$

Proof. According to (7), one has

$$V(X_k) = E\left[X_k^{\mathrm{T}}Q_{1i}X_k + u_k^{\mathrm{T}}R_iu_k + \sum_{z=k+1}^{\infty}(X_z^{\mathrm{T}}Q_{1i}X_z + u_z^{\mathrm{T}}R_iu_z)\right], \tag{9}$$

which yields

$$V(X_k) = E\left[X_k^{\mathrm{T}}Q_{1i}X_k + u_k^{\mathrm{T}}R_iu_k\right] + V(X_{k+1}). \tag{10}$$

Substituting (7) into (10), we obtain the Bellman equation as

$$X_k^{\mathrm{T}}P_iX_k = X_k^{\mathrm{T}}Q_{1i}X_k + u_k^{\mathrm{T}}R_iu_k + X_{k+1}^{\mathrm{T}}\left(\sum_{j=1}^{m}\pi_{ij}P_j\right)X_{k+1}. \tag{11}$$

Next, the augmented Hamiltonian function can be defined as

$$H(X_k, u_k) = X_k^{\mathrm{T}}Q_{1i}X_k + u_k^{\mathrm{T}}R_iu_k + X_{k+1}^{\mathrm{T}}\left(\sum_{j=1}^{m}\pi_{ij}P_j\right)X_{k+1} - X_k^{\mathrm{T}}P_iX_k. \tag{12}$$

Applying the stationary condition according to [36], we have

$$\frac{\partial H(X_k, u_k)}{\partial u_k} = 2R_iu_k + 2B_{1i}^{\mathrm{T}}\left(\sum_{j=1}^{m}\pi_{ij}P_j\right)X_{k+1} = 0. \tag{13}$$

To this end, the optimal controller is obtained as

$$\begin{aligned} u_k^* =& -K_i^*X_k \\ =& -\left[R_i + B_{1i}^{\mathrm{T}}\left(\sum_{j=1}^{m}\pi_{ij}P_j\right)B_{1i}\right]^{-1}B_{1i}^{\mathrm{T}}\left(\sum_{j=1}^{m}\pi_{ij}P_j\right)T_iX_k. \end{aligned} \tag{14}$$

Substituting (14) into (11), we can derive (8). This completes the proof. □

3. Main results

In this section, two algorithms are developed to address the OTC problem for MJLSs. Firstly, the PI technique is applied to derive a unique solution of the ACGARE in (8). Subsequently, an online optimization algorithm for solving the OTC problem is summarized by assuming the dynamic knowledge of the system without requiring the transition probability in subsection 3.1. To eliminate the usage of system dynamics, a fully model-free off-policy RL algorithm is further developed based on the analysis of the online optimization algorithm in subsection 3.2.

Inspired by the PI technique, the ACGARE in (8) can be iteratively solved by the following theorem.

Theorem 2. *The solution matrix $P_i^{\eta+1}$ generated by the following two-step iteration equations (15) and (16), which converge to the unique positive-definite solution of (8), that is, $lim_{k\to\infty}P_i^{\eta}=P_i^{*}$.*

$$X_k^{\mathrm{T}}P_i^{\eta+1}X_k = X_k^{\mathrm{T}}Q_{1i}X_k + (u_k^{\eta})^{\mathrm{T}}R_iu_k^{\eta} + X_{k+1}^{\mathrm{T}}\times\left(\sum_{j=1}^{m}\pi_{ij}P_j^{\eta+1}\right)X_{k+1}, \tag{15}$$

$$K_i^{\eta+1} = -\left[R_i + B_{1i}^{\mathrm{T}}\left(\sum_{j=1}^{m}\pi_{ij}P_j^{\eta}\right)B_{1i}\right]^{-1}\times B_{1i}^{\mathrm{T}}\left(\sum_{j=1}^{m}\pi_{ij}P_j^{\eta}\right)T_i. \tag{16}$$

Proof. Substituting the control policy $u_k^{\eta}=-K_i^{\eta}X_k$ into the Bellman Eq. (15), we have

$$\begin{aligned}X_k^{\mathrm{T}}P_i^{\eta+1}X_k =& X_k^{\mathrm{T}}T_i^{\mathrm{T}}\left(\sum_{j=1}^{m}\pi_{ij}P_j^{\eta+1}\right)T_iX_k + X_k^{\mathrm{T}}Q_{1i}X_k\\ &- X_k^{\mathrm{T}}T_i^{\mathrm{T}}\left(\sum_{j=1}^{m}\pi_{ij}P_j^{\eta}\right)B_{1i}\left[B_{1i}^{\mathrm{T}}\left(\sum_{j=1}^{m}\pi_{ij}P_j^{\eta}\right)B_{1i}+R_i\right]^{-1}B_{1i}^{\mathrm{T}}\left(\sum_{j=1}^{m}\pi_{ij}P_j^{\eta}\right)T_iX_k,\end{aligned} \tag{17}$$

Recalling to the results of the monotonic convergence of positive operators [37], we take limit on both sides of (17) and consider that $lim_{k\to\infty}P_i^{\eta}=lim_{k\to\infty}P_i^{\eta+1}=P_i^{\infty}$. Then, we can get

$$\begin{aligned}X_k^{\mathrm{T}}P_i^{\infty}X_k =& X_k^{\mathrm{T}}T_i^{\mathrm{T}}\left(\sum_{j=1}^{m}\pi_{ij}P_j^{\infty}\right)T_iX_k + X_k^{\mathrm{T}}Q_{1i}X_k - X_k^{\mathrm{T}}T_i^{\mathrm{T}}\\ &\times\left(\sum_{j=1}^{m}\pi_{ij}P_j^{\infty}\right)B_{1i}\left[B_{1i}^{\mathrm{T}}(\sum_{j=1}^{m}\pi_{ij}P_j^{\infty})B_{1i}+R_i\right]^{-1}\times B_{1i}^{\mathrm{T}}\left(\sum_{j=1}^{m}\pi_{ij}P_j^{\infty}\right)T_iX_k.\end{aligned} \tag{18}$$

Simplifying (18), we have

$$\begin{aligned}P_i^{\infty} =& T_i^{\mathrm{T}}\left(\sum_{j=1}^{m}\pi_{ij}P_j^{\infty}\right)T_i + Q_{1i} - T_i^{\mathrm{T}}\left(\sum_{j=1}^{m}\pi_{ij}P_j^{\infty}\right)B_{1i}\left[B_{1i}^{\mathrm{T}}\right.\\ &\left.\times\left(\sum_{j=1}^{m}\pi_{ij}P_j^{\infty}\right)B_{1i}+R_i\right]^{-1}B_{1i}^{\mathrm{T}}\left(\sum_{j=1}^{m}\pi_{ij}P_j^{\infty}\right)T_i.\end{aligned} \tag{19}$$

It can be easily seen that Eqs. (8) and (19) have the same form. Since P_i^* is the unique solution of (8), we have $lim_{k\to\infty}P_i^\eta = P_i^*$. The proof is completed. □

3.1. Online optimization algorithm

In this section, an online RL method is implemented to address the OTC problem.

Based on Kronecker production, we have

$$\varkappa^T W \varrho = (\varkappa^T \otimes \varrho^T) vec(W), \tag{20}$$

where $\varkappa \in \Re^t$, $\varrho \in \Re^z$ and $W \in \Re^{t\times z}$. $vec(\cdot)$ is defined as a function, which stacks the columns into a single vector.

The Bellman Eq. (15) can be rewritten as

$$\begin{aligned}&(X_k^T \otimes X_k^T) vec(P_i^{\eta+1}) - (X_{k+1}^T \otimes X_{k+1}^T) vec\left(\sum_{j=1}^{m} \pi_{ij} P_j^{\eta+1}\right)\\ &= X_k^{\mathrm{T}} Q_{1i} X_k + (u_k^\eta)^{\mathrm{T}} R_i u_k^\eta,\end{aligned} \tag{21}$$

Using least-square (LS) method, $P_i^{\eta+1}$and $K_i^{\eta+1}$ can be obtained simultaneously. For the positive integer $s \geq n^2$, one defines

$$\Phi_i^\eta = \begin{bmatrix} X_k^T Q_{1i} X_k + X_k^T (K_i^\eta)^T R_i K_i^\eta X_k \\ X_{k+1}^T Q_{1i} X_{k+1} + X_{k+1}^T (K_i^\eta)^T R_i K_i^\eta X_{k+1} \\ \vdots \\ X_{k+s-1}^T Q_{1i} X_{k+s-1} + X_{k+s-1}^T (K_i^\eta)^T R_i K_i^\eta X_{k+s-1} \end{bmatrix}, \tag{22}$$

$$\Psi_i^\eta = \begin{bmatrix} \Delta_{(1)XX} & \Delta_{(1)XX^1} \\ \Delta_{(2)XX} & \Delta_{(2)XX^1} \\ \vdots & \vdots \\ \Delta_{(s)XX} & \Delta_{(s)XX^1} \end{bmatrix}, \tag{23}$$

where

$$\Delta_{(a)XX} = (X_{k+a-1}^T \otimes X_{k+a-1}^T),$$
$$\Delta_{(a)XX^1} = -(X_{k+a}^T \otimes X_{k+a}^T), a = (1, 2, \cdots, s).$$

Define the unknown matrices in the Bellman Eq. (21) as

$$\Omega_i^\eta = \left[vec(\omega_{1i}^{\eta+1})^T \quad vec(\omega_{2i}^{\eta+1})^T\right]^T, \tag{24}$$

where $\omega_{1i}^{\eta+1} = P_i^{\eta+1}$, $\omega_{2i}^{\eta+1} = \sum_{j=1}^m \pi_{ij} P_j^{\eta+1}$.

Based on (21)-(24), one has

$$\Psi_i^\eta \Omega_i^\eta = \Phi_i^\eta. \tag{25}$$

To this end, (25) can be solved by the LS method shown as

$$\Omega_i^\eta = \left[(\Psi_i^\eta)^T \Psi_i^\eta\right]^{-1} (\Psi_i^\eta)^T \Phi_i^\eta. \tag{26}$$

Using Ω_i^{η}, the gain $K_i^{\eta+1}$ can be obtained as

$$K_i^{\eta+1} = \left(R_i + B_{1i}^{\mathrm{T}}\omega_{2i}^{\eta+1}B_{1i}\right)^{-1}B_{1i}^{\mathrm{T}}\omega_{2i}^{\eta+1}T_i. \tag{27}$$

The online learning algorithm is summarized in Algorithm 1. Notice that Algorithm 1 can

Algorithm 1: Online PI algorithm for discrete-time MJLSs.

Input: Give the stabilizing control policies $u_i^0 = -K_i^0 X_k$ and an arbitrarily small constant ϵ.

Output: The optimal control policy K_i^*.

$\eta \leftarrow 0$;
while $\eta \geqslant 0$ **do**
  for $i \leftarrow 1$ ***to*** m **do**
    Solve for

$$\Omega_i^{\eta} = \left[(\Psi_i^{\eta})^T \Psi_i^{\eta}\right]^{-1}(\Psi_i^{\eta})^T \Phi_i^{\eta}.$$

    Calculate $P_i^{\eta+1}$ and update $K_i^{\eta+1}$

$$P_i^{\eta+1} = \omega_{1i}^{\eta+1},$$
$$K_i^{\eta+1} = \left(R_i + B_{1i}^T\omega_{2i}^{\eta+1}B_{1i}\right)^{-1}B_{1i}^T\omega_{2i}^{\eta+1}T_i.$$

    if $\max \|P_i^{\eta+1} - P_i^{\eta}\| \geq \epsilon$ **then**
      $\eta \leftarrow \eta + 1$;
    else
      $K_i^* = K_i^{\eta+1}$;
      break ;
Return K_i^*.

obtain the optimal control policies without using the knowledge of transition probability; but it requires the augmented system dynamics (T_i, B_{1i}). To obviate the requirement for system dynamics, a fully model-free approach is proposed to address the OTC problem in next subsection.

3.2. Model-free off-policy RL algorithm

Based on Algorithm 1, this section presents a novel model-free algorithm. In this algorithm, we assume the optimal control policies cannot get access to the system dynamics and transition probability. In this algorithm, we rewrite the system (6) as

$$X_{k+1} = (T_i - B_{1i}K_i^{\eta})X_k + B_{1i}(u_k + K_i^{\eta}X_k). \tag{28}$$

Based on (28), we have

$$X_k^T P_i^{\eta+1} X_k - X_{k+1}^T\left(\sum_{j=1}^{m}\pi_{ij}P_j^{\eta+1}\right)X_{k+1}$$

$$
\begin{aligned}
=&X_k^T P_i^{\eta+1} X_k - X_k^T (T_i - B_{1i}K_i^\eta)^T \left(\sum_{j=1}^{m} \pi_{ij} P_j^{\eta+1}\right)(T_i - B_{1i}K_i^\eta)X_k \\
&- (u_k + K_i^\eta X_k)^T B_{1i}^T \left(\sum_{j=1}^{m} \pi_{ij} P_j^{\eta+1}\right)(T_i - B_{1i}K_i^\eta)X_k \\
&- (u_k + K_i^\eta X_k)^T B_{1i}^T \left(\sum_{j=1}^{m} \pi_{ij} P_j^{\eta+1}\right)X_{k+1}.
\end{aligned} \tag{29}
$$

Then, we derive the following Lyapunov equation by (15)

$$
0 = Q_{1i} + (K_i^\eta)^T R_i K_i^\eta - P_i^{\eta+1} + (T_i - B_{1i}K_i^\eta)^T \left(\sum_{j=1}^{m} \pi_{ij} P_j^{\eta+1}\right)(T_i - B_{1i}K_i^\eta). \tag{30}
$$

Based on (30), Eq. (29) can be rewritten as

$$
\begin{aligned}
&X_k^T P_i^{\eta+1} X_k - X_{k+1}^T \left(\sum_{j=1}^{m} \pi_{ij} P_j^{\eta+1}\right)X_{k+1} \\
=&X_k^T Q_{1i} X_k + X_k^T (K_i^\eta)^T R_i K_i^\eta X_k - (u_k + K_i^\eta X_k)^T B_{1i}^T \left(\sum_{j=1}^{m} \pi_{ij} P_j^{\eta+1}\right) \\
&\times (T_i - B_{1i}K_i^\eta)X_k - (u_k + K_i^\eta X_k)^T B_{1i}^T \left(\sum_{j=1}^{m} \pi_{ij} P_j^{\eta+1}\right)X_{k+1} \\
=&X_k^T Q_{1i} X_k + X_k^T (K_i^\eta)^T R_i K_i^\eta X_k - (u_k + K_i^\eta X_k)^T B_{1i}^T \\
&\times \left(\sum_{j=1}^{m} \pi_{ij} P_j^{\eta+1}\right)T_i X_k + (u_k + K_i^\eta X_k)^T B_{1i}^T \\
&\times \left(\sum_{j=1}^{m} \pi_{ij} P_j^{\eta+1}\right)B_{1i}K_i^\eta X_k - (u_k + K_i^\eta X_k)^T B_{1i}^T \\
&\times \left(\sum_{j=1}^{m} \pi_{ij} P_j^{\eta+1}\right)T_i X_k + (u_k + K_i^\eta X_k)^T B_{1i}^T \\
&\times \left(\sum_{j=1}^{m} \pi_{ij} P_j^{\eta+1}\right)B_{1i}K_i^\eta X_k - (u_k + K_i^\eta X_k)^T B_{1i}^T \\
&\times \left(\sum_{j=1}^{m} \pi_{ij} P_j^{\eta+1}\right)B_{1i}u_k - (u_k + K_i^\eta X_k)^T B_{1i}^T \left(\sum_{j=1}^{m} \pi_{ij} P_j^{\eta+1}\right)B_{1i}K_i^\eta X_k \\
=&X_k^T Q_{1i} X_k + X_k^T (K_i^\eta)^T R_i K_i^\eta X_k - 2(u_k + K_i^\eta X_k)^T B_{1i}^T \\
&\times \left(\sum_{j=1}^{m} \pi_{ij} P_j^{\eta+1}\right)T_i X_k + (u_k + K_i^\eta X_k)^T B_{1i}^T \\
&\times \left(\sum_{j=1}^{m} \pi_{ij} P_j^{\eta+1}\right)B_{1i}(K_i^\eta X_k - u_k)
\end{aligned} \tag{31}
$$

Similar to (20), Eq. (31) can be represented as

$$
\begin{aligned}
&(X_k^T \otimes X_k^T)vec(P_i^{\eta+1}) - (X_{k+1}^T \otimes X_{k+1}^T)vec\left(\sum_{j=1}^{m}\pi_{ij}P_j^{\eta+1}\right)\\
&+2\left[X_k^T \otimes (u_k + K_i^{\eta}X_k)^T\right]vec\left[B_{1i}^T\left(\sum_{j=1}^{m}\pi_{ij}P_j^{\eta+1}\right)T_i\right] - \left[(K_i^{\eta}\right.\\
&\left.\times X_k - u_k\right)^T \otimes (K_i^{\eta}X_k + u_k)^T\Bigg]vec\left[B_{1i}^T\left(\sum_{j=1}^{m}\pi_{ij}P_j^{\eta+1}\right)B_{1i}\right]\\
=&X_k^T Q_{1i}X_k + X_k^T(K_i^{\eta})^T R_i K_i^{\eta}X_k. \qquad (32)
\end{aligned}
$$

Using LS method, $P_i^{\eta+1}$and $K_i^{\eta+1}$ can be obtained simultaneously. The Bellman Eq. (32) has $2n^2+mn+m^2$ unknown parameters. Therefore, at least $2n^2+mn+m^2$ data sets are provided before Eq. (32) can be solved by using LS method. For a certain positive integer $l(l \geq 2n^2+mn+m^2)$, we define the following equalities

$$
\Xi_i^{\eta} = \begin{bmatrix} X_k^T Q_{1i}X_k + X_k^T(K_i^{\eta})^T R_i K_i^{\eta}X_k \\ X_{k+1}^T Q_{1i}X_{k+1} + X_{k+1}^T(K_i^{\eta})^T R_i K_i^{\eta}X_{k+1} \\ \vdots \\ X_{k+l-1}^T Q_{1i}X_{k+l-1} + X_{k+l-1}^T(K_i^{\eta})^T R_i K_i^{\eta}X_{k+l-1} \end{bmatrix}, \qquad (33)
$$

$$
\beth_i^{\eta} = \begin{bmatrix} \Lambda_{(1)XX} & \Lambda_{(1)XX^1} & \Lambda_{(1)Xu} & \Lambda_{(1)uu} \\ \Lambda_{(2)XX} & \Lambda_{(2)XX^1} & \Lambda_{(2)Xu} & \Lambda_{(2)uu} \\ \vdots & \vdots & \vdots & \vdots \\ \Lambda_{(l)XX} & \Lambda_{(l)XX^1} & \Lambda_{(l)Xu} & \Lambda_{(l)uu} \end{bmatrix}, \qquad (34)
$$

where

$$
\begin{aligned}
&\Lambda_{(a)XX} = (X_{k+a-1}^T \otimes X_{k+a-1}^T),\\
&\Lambda_{(a)XX^1} = -(X_{k+a}^T \otimes X_{k+a}^T),\\
&\Lambda_{(a)Xu} = 2\left[X_{k+a-1}^T \otimes (K_i^{\eta}X_{k+a-1} + u_{k+a-1})^T\right],\\
&\Lambda_{(a)uu} = -\left[(K_i^{\eta}X_{k+a-1} - u_{k+a-1})^T \otimes (K_i^{\eta}X_{k+a-1} + u_{k+a-1})^T\right], a = (1, 2, \cdots, l).
\end{aligned}
$$

Define the unknown matrices in equation(32) as

$$
\Theta_i^{\eta} = \left[vec(\theta_{1i}^{\eta+1})^T \quad vec(\theta_{2i}^{\eta+1})^T \quad vec(\theta_{3i}^{\eta+1})^T \quad vec(\theta_{4i}^{\eta+1})^T\right]^T, \qquad (35)
$$

where

$$
\begin{aligned}
&\theta_{1i}^{\eta+1} = P_i^{\eta+1},\\
&\theta_{2i}^{\eta+1} = \sum_{j=1}^{m}\pi_{ij}P_j^{\eta+1},\\
&\theta_{3i}^{\eta+1} = B_{1i}^T\left(\sum_{j=1}^{m}\pi_{ij}P_j^{\eta+1}\right)T_i,
\end{aligned}
$$

$$\theta_{4i}^{\eta+1} = B_{1i}^T\left(\sum_{j=1}^{m}\pi_{ij}P_j^{\eta+1}\right)B_{1i}.$$

Based on (32)-(35), one has

$$\beth_i^\eta \Theta_i^\eta = \Xi_i^\eta, \tag{36}$$

Considering the form of (36), we use LS method to solve it in Theorem 3.

Theorem 3. *If there exists a positive $l_0 > 0$, the following condition satisfies $\forall l \geq l_0$:*

$$rank(\beth_i^\eta) = 2n^2 + nm + m^2. \tag{37}$$

If $\beth_i^\eta$ have column full rank, (36) can be directly presented as

$$\Theta_i^\eta = \left[(\beth_i^\eta)^T\beth_i^\eta\right]^{-1}(\beth_i^\eta)^T\Xi_i^\eta. \tag{38}$$

Proof. Assuming that (36) has two different solutions Θ_{1i}^η and Θ_{2i}^η, we have $\beth_i^\eta\Theta_{1i}^\eta = \Xi_i^\eta$ and $\beth_i^\eta\Theta_{2i}^\eta = \Xi_i^\eta$. Subtracting two matrix equations, we have $\beth_i^\eta(\Theta_{1i}^\eta - \Theta_{2i}^\eta) = 0$. Notice from (37) that $\beth_i^\eta$ has column full rank, $\beth_i^\eta(\Theta_{1i}^\eta - \Theta_{2i}^\eta) = 0$ has the unique solution $\Theta_{1i}^\eta - \Theta_{2i}^\eta = 0$ (or $\Theta_{1i}^\eta = \Theta_{2i}^\eta$). However, this result contradicts our assumption. Therefore, (19) has the unique solution by solving (38). This completes the proof. □

Using Θ_i^η, the gain $K_i^{\eta+1}$ can be obtained by:

$$K_i^{\eta+1} = \left(R_i + \theta_{4i}^{\eta+1}\right)^{-1}\theta_{3i}^{\eta+1}. \tag{39}$$

Theorem 4. *Given the initial stabilizing values K_i^0, when the condition of Theorem 2 and Theorem 3 hold, the solutions of (38) are equivalent to the optimal solutions of (8).*

Proof. Set the initial stabilizing values K_i^0. According to Theorem 2, if there exists a positive-definite solution $P_i^\eta = (P_i^\eta)^T$ satisfying (8), K_i^η can be obtained by the policy iteration in Algorithm 2. Then, (31) is deduced if the solution of (28) is solved by (15) and (16). Besides, (31) accords with (38) based on Theorem 3. Thus, the solutions of Eq. (38) are equivalent to the optimal solutions of (8). This completes the proof. □

Remark 3. By Theorem 3 and Theorem 4, the proposed Algorithm 2 can iteratively solve (38) without involving any knowledge of system dynamics and transition probability. Algorithm 2 contains two steps. In the first step, in (38), the gains $\theta_{1i}^{\eta+1}$ through $\theta_{4i}^{\eta+1}$ are calculated by using the measured data Ξ_i^η and $\beth_i^\eta$ (see equalities (33) and (34)). In the second step, the control strategy is updated by using the gains learned in the first step. Moreover, this algorithm allows us to conduct in-depth studies of other hybrid systems[38–41] or nonlinear systems[42–45].

4. Numerical example

In this part, a simulation example is shown to verify the property and validity of the proposed algorithm. Here, the Markov process $\{\tau(k), k \geq 0\}$ used in the simulation consists of two modes of MJSs. The dynamics of these systems are set as follows:

Algorithm 2: Model-free off-policy algorithm for discrete-time MJLSs.

Input: Give the stabilizing control policies $u_i^0 = -K_i^0 X_k + \hat{e}$ ($\hat{e}$ is the explosion signal employed in $[0, \eta_{\hat{e}}]$) and an arbitrarily small constant ϵ.

Output: The optimal control policy K_i^*.

$\eta \leftarrow 0$;
while $\eta \geqslant 0$ **do**
  for $i \leftarrow 1$ ***to*** m **do**
    if $\eta > \eta_{\hat{e}}$ **then**
      $\hat{e} = 0$;
    Solve for
    $\Theta_i^\eta = \left[(\beth_i^\eta)^T \beth_i^\eta\right]^{-1} (\beth_i^\eta)^T \Xi_i^\eta$.
    Calculate $P_i^{\eta+1}$ and update $K_i^{\eta+1}$
    $P_i^{\eta+1} = \theta_{1i}^{\eta+1}$,
    $K_i^{\eta+1} = \left(R_i + \theta_{4i}^{\eta+1}\right)^{-1} \theta_{3i}^{\eta+1}$.
    if $\max \|P_i^{\eta+1} - P_i^\eta\| \geq \epsilon$ **then**
      $\eta \leftarrow \eta + 1$;
    else
      $K_i^* = K_i^{\eta+1}$;
      break ;
Return K_i^*.

Mode 1:

$$A_1 = \begin{bmatrix} 1.00000 & 0.10000 \\ -0.03270 & 0.80000 \end{bmatrix}, B_1 = \begin{bmatrix} 0 \\ 0.03333 \end{bmatrix},$$
$$C_1 = \begin{bmatrix} 1.00000 & 1.00000 \end{bmatrix}, R_1 = 1, Q_1 = 1.$$

Mode 2:

$$A_2 = \begin{bmatrix} 1.00000 & 0.10000 \\ -0.07358 & 0.85000 \end{bmatrix}, B_2 = \begin{bmatrix} 0 \\ 0.02500 \end{bmatrix},$$
$$C_2 = \begin{bmatrix} 1.00000 & 1.00000 \end{bmatrix}, R_2 = 1, Q_2 = 1.$$

Regarding the reference signal $r_{k+1} = Fr_k$, we set $F = 0.94$. The initial states of the augmented system are given as $X_1^0 = X_2^0 = \begin{bmatrix} 1 & -1 & 1 \end{bmatrix}^T$ and the transition probability matrix is:

$$\Pi = \begin{bmatrix} 0.3 & 0.7 \\ 0.7 & 0.3 \end{bmatrix}.$$

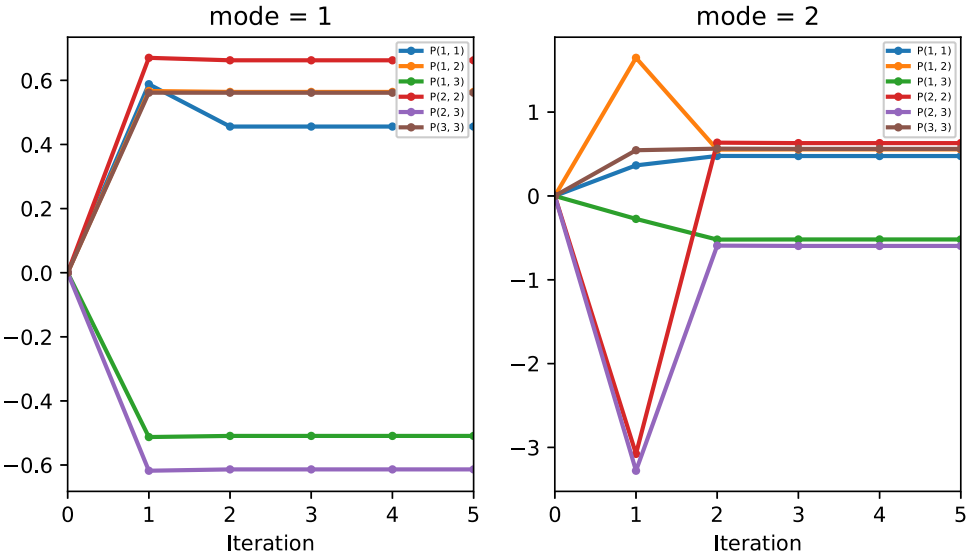

Fig. 1. The optimal values of P_i^{off} and the convergence.

Firstly, the online learning Algorithm 1 is implemented to solve the ACGARE, and the simulation results show that the values of P_i finally converge to

$$P_1 = \begin{bmatrix} 0.45593156 & 0.56378728 & -0.50928221 \\ 0.56378728 & 0.66240760 & -0.61362200 \\ -0.50928221 & -0.61362200 & 0.56145152 \end{bmatrix},$$

$$P_2 = \begin{bmatrix} 0.47604179 & 0.55489969 & -0.51905281 \\ 0.55489969 & 0.63010383 & -0.59638985 \\ -0.51905281 & -0.59638985 & 0.56150113 \end{bmatrix},$$

and the control gains K_i converge to

$$K_1 = \begin{bmatrix} -0.01584223 & -0.01219772 & 0.01399581 \end{bmatrix},$$
$$K_2 = \begin{bmatrix} -0.01145101 & -0.00948805 & 0.01036660 \end{bmatrix}.$$

Secondly, the model-free off-policy RL Algorithm 2 is implemented to solve the ACGARE. Select $u_i^0 = -K_i^0 X_k + \hat{e}$ and set $\hat{e} = 0.1\hat{e}^{-0.01t}\sin(30t) + 0.1\sin(10t)$. The values of P_i^{off} matrices converge to

$$P_1^{off} = \begin{bmatrix} 0.45591266 & 0.56391543 & -0.50927506 \\ 0.56391543 & 0.66256976 & -0.61372715 \\ -0.50927506 & -0.61372715 & 0.56145159 \end{bmatrix},$$

$$P_2^{off} = \begin{bmatrix} 0.47602836 & 0.55496709 & -0.51904733 \\ 0.55496709 & 0.63020357 & -0.59644908 \\ -0.51904733 & -0.59644908 & 0.56150115 \end{bmatrix},$$

and the control gains K_i^{off} converge to

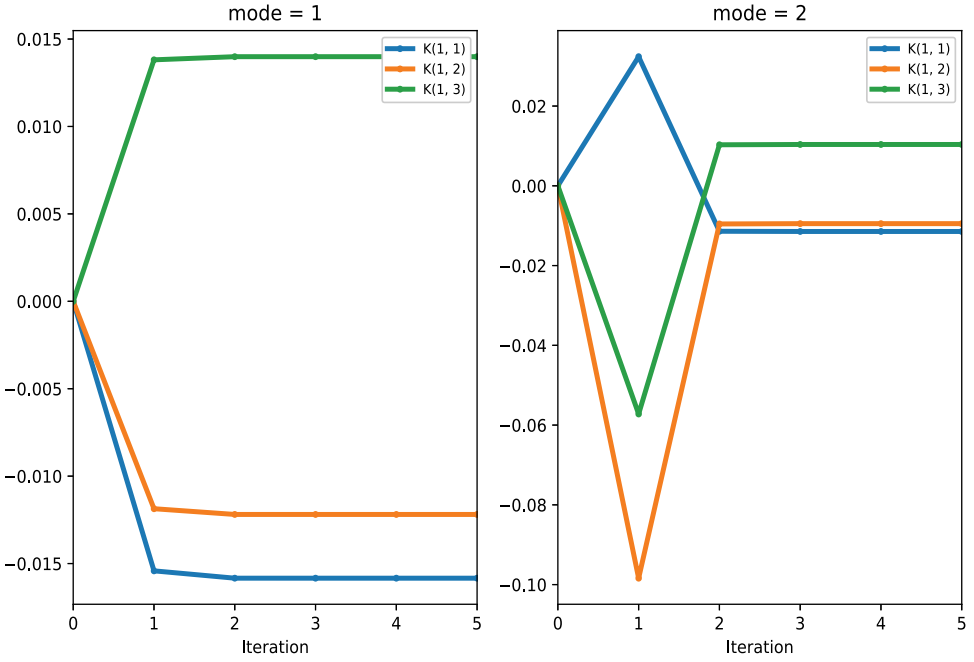

Fig. 2. The optimal values of K_i^{off} and the convergence.

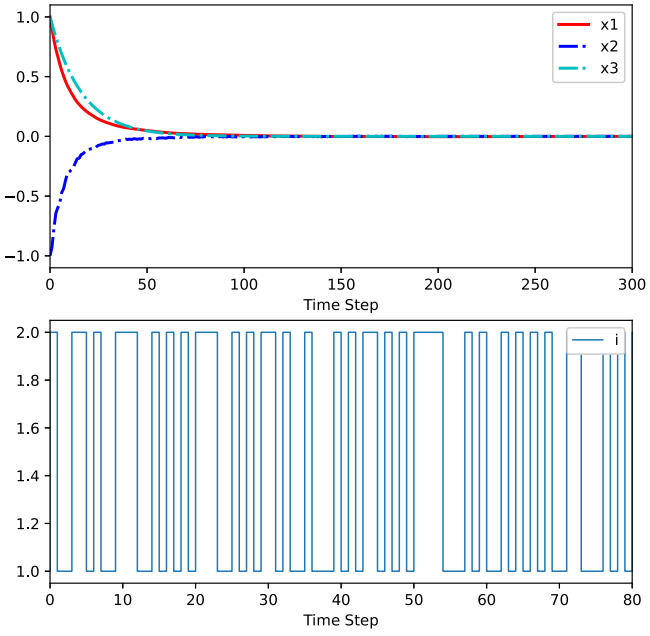

Fig. 3. The system states and the jumping modes.

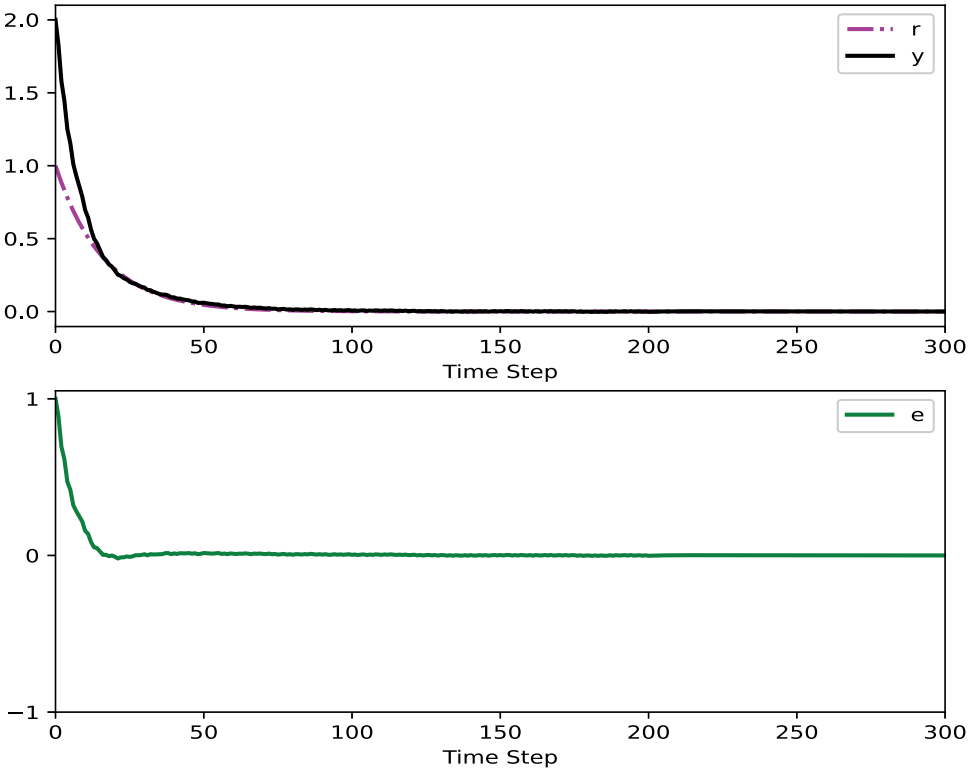

Fig. 4. The output tracks reference trajectories and the tracking errors.

$$K_1^{off} = \begin{bmatrix} -0.01583684 & -0.01219152 & 0.01399142 \end{bmatrix},$$
$$K_2^{off} = \begin{bmatrix} -0.01144900 & -0.00948534 & 0.01036485 \end{bmatrix}.$$

Algorithm 1 is implemented to solve the ACGARE, which requires the knowledge of the system dynamics (T_i, B_{1i}), whereas Algorithm 2 also achieves almost the same solution of the ACGARE with completely unknown dynamics and transition probability. It seems that Algorithm 2 is more efficient than Algorithm 1 at the same number of iteration steps. As in Algorithm 2, Fig. 1 shows that the parameters of P_i^{off} converge to the stable value after two iterations. Fig. 2 shows that the control gain K_i converge to the stable value after two iterations. Fig. 3 shows the system states and the jumping modes, where all states are shown to be stabilized and converged to zero by the designed controller. In Fig. 4, the output tracks the reference signal and the tracking error tends to zero. Especially at 0.3s, the output gradually tracks the reference trajectory; when the number of iteration steps increases, the two curves converge to a stable value.

5. Conclusion

In this study, a fully model-free off-policy RL algorithm has been proposed to address the OTC problem for discrete-time MJLSs. By constructing the augmented system, we design the off-policy RL algorithm to solve the ACGARE. The advantage is that we do not need any

knowledge of system dynamics and transition probability. The simulation results verify the convergence of the proposed algorithm, which show that the given reference signal has been well-tracked by the optimal controller iteratively derived from the proposed algorithm.

Declaration of Competing Interest

The authors declare that they have no conflicts of interest to this work. We still declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (No. 62073001), the Anhui Provincial Key Research and Development Project (No. 2022i01020013), and the University Synergy Innovation Program of Anhui Province (No. GXXT-2021-010).

References

[1] B. Jiang, H. R. Karimi, Y. Kao, et al., Takagi-sugeno model based event-triggered fuzzy sliding-mode control of networked control systems with semi-markovian switchings, 2019, IEEE Trans. Fuzzy Syst., 28, 4, 673–683
[2] C. Zhao, J. Jiang, Y. Guan, et al., EMR-Based medical knowledge representation and inference via markov random fields and distributed representation learning, Artif. Intell. Med. 87 (2018) 49–59.
[3] W. Qi, G. Zong, W.X. Zheng, Adaptive event-triggered SMC for stochastic switching systems with semi-markov process and application to boost converter circuit model, IEEE Trans. Circuits Syst. I: Regul. Pap. 68 (2) (2020) 786–796.
[4] O.L.V. Costa, M.D. Fragoso, R.P. Marques, D.-t. M. j. l. systems, Springer Science & Business Media (2006).
[5] M. Zhang, C. Shen, Z.G. Wu, Asynchronous observer-based control for exponential stabilization of markov jump systems, IEEE Trans. Circuits Syst. II: Express Briefs 67 (10) (2019a) 2039–2043.
[6] F. Li, S. Song, J. Zhao, et al., Synchronization control for markov jump neural networks subject to HMM observation and partially known detection probabilities, Appl. Math. Comput. 360 (2019) 1–13.
[7] G. Ran, J. Liu, C. Li, et al., Fuzzy-model-based asynchronous fault detection for markov jump systems with partially unknown transition probabilities: an adaptive event-triggered approach, IEEE Trans. Fuzzy Syst. (2022).
[8] Y. Zhang, C.C. Lim, F. Liu, Robust mixed h_2/h_∞ model predictive control for markov jump systems with partially uncertain transition probabilities, J. Franklin Inst. 355 (8) (2018) 3423–3437.
[9] G. Zong, D. Yang, L. Hou, et al., Robust finite-time h_∞ control for markovian jump systems with partially known transition probabilities, J. Franklin Inst. 350 (6) (2013) 1562–1578.
[10] Y. Ma, X. Jia, D. Liu, Finite-time dissipative control for singular discrete-time markovian jump systems with actuator saturation and partly unknown transition rates, Appl. Math. Model. 53 (2018) 49–70.
[11] R. Li, J. Cao, Finite-time stability analysis for markovian jump memristive neural networks with partly unknown transition probabilities, IEEE Trans. Neural Netw. Learn. Syst. 28 (12) (2016) 2924–2935.
[12] J. Xia, J.H. Park, B. Zhang, et al., Robust h_∞ tracking control for uncertain markovian jumping systems with interval time-varying delay, Complexity 21 (2) (2015) 355–366.
[13] G. Zhuang, J. Xia, J. Zhao, et al., Nonfragile h_∞ output tracking control for uncertain singular markovian jump delay systems with network-induced delays and data packet dropouts, Complexity 21 (6) (2016) 396–411.
[14] H. Zhang, K. Zhang, Y. Cai, et al., Adaptive fuzzy fault-tolerant tracking control for partially unknown systems with actuator faults via integral reinforcement learning method, IEEE Trans. Fuzzy Syst. 27 (10) (2019) 1986–1998.
[15] K. Shojaei, Three-dimensional neural network tracking control of a moving target by underactuated autonomous underwater vehicles, Neural Comput. Appl. 31 (2) (2019) 509–521.
[16] Z. Ni, H. He, J. Wen, Adaptive learning in tracking control based on the dual critic network design, IEEE Trans. Neural Netw. Learn. Syst. 24 (6) (2013) 913–928.

[17] C. Qin, H. Zhang, Y. Luo, Optimal tracking control of a class of nonlinear discrete-time switched systems using adaptive dynamic programming, Neural Comput. Appl. 24 (3) (2014) 531–538.
[18] Q. Wei, D. Liu, Adaptive dynamic programming for optimal tracking control of unknown nonlinear systems with application to coal gasification, IEEE Trans. Autom. Sci. Eng. 11 (4) (2013) 1020–1036.
[19] Y. Huang, D. Liu, Neural-network-based optimal tracking control scheme for a class of unknown discrete-time nonlinear systems using iterative ADP algorithm, Neurocomputing 125 (2014) 46–56.
[20] M. Zhang, Z. Wu, J. Yan, et al., Attack-resilient optimal PMU placement via reinforcement learning guided tree search in smart grids, IEEE Trans. Inf. Forensics Secur. 17 (2022) 1919–1929.
[21] H. Modares, F.L. Lewis, Optimal tracking control of nonlinear partially-unknown constrained-input systems using integral reinforcement learning, Automatica 50 (7) (2014) 1780–1792.
[22] H. Zhang, Q. Wei, Y. Luo, 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. Part B (Cybern.) 38 (4) (2008) 937–942.
[23] Y.M. Park, M.S. . Choi, K.Y. Lee, An optimal tracking neuro-controller for nonlinear dynamic systems, IEEE Trans. Neural Netw. 7 (5) (1996) 1099–1110.
[24] C. Qin, H. Zhang, Y. Luo, Online optimal tracking control of continuous-time linear systems with unknown dynamics by using adaptive dynamic programming, Int. J. Control 87 (5) (2014) 1000–1009.
[25] Y. Liu, H. Zhang, R. Yu, et al., Data-driven optimal tracking control for discrete-time systems with delays using adaptive dynamic programming, J. Franklin Inst. 355 (13) (2018) 5649–5666.
[26] D. Liu, Q. Wei, Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems, IEEE Trans. Neural Netw. Learn. Syst. 25 (3) (2013) 621–634.
[27] D. Vrabie, O. Pastravanu, M. Abu-Khalaf, et al., Adaptive optimal control for continuous-time linear systems based on policy iteration, Automatica 45 (2) (2009) 477–484.
[28] S. He, H. Fang, M. Zhang, et al., Adaptive optimal control for a class of nonlinear systems: the online policy iteration approach, IEEE Trans. Neural Netw. Learn. Syst. 31 (2) (2019) 549–558.
[29] B. Kiumarsi, F.L. Lewis, Actor-critic-based optimal tracking for partially unknown nonlinear discrete-time systems, IEEE Trans. Neural Netw. Learn. Syst. 26 (1) (2014) 140–151.
[30] H. Zhang, L. Cui, X. Zhang, et al., Data-driven robust approximate optimal tracking control for unknown general nonlinear systems using adaptive dynamic programming method, IEEE Trans. Neural Netw. 22 (12) (2011) 2226–2236.
[31] K. Zhang, H. Zhang, Y. . Cai, et al., Parallel optimal tracking control schemes for mode-dependent control of coupled markov jump systems via integral RL method, IEEE Trans. Autom. Sci. Eng. 17 (3) (2019) 1332–1342.
[32] H. Jiang, H. Zhang, Y. Luo, et al., Optimal tracking control for completely unknown nonlinear discrete-time markov jump systems using data-based reinforcement learning method, Neurocomputing 194 (2016) 176–182.
[33] B. Kiumarsi, F.L. Lewis, Z.P. Jiang, h_∞Control of linear discrete-time systems: off-policy reinforcement learning, Automatica 78 (2017) 144–152.
[34] Y. Jiang, Z.P. Jiang, Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics, Automatica 48 (10) (2012) 2699–2704.
[35] Y. Wen, H. Zhang, H. Su, et al., Optimal tracking control for non-zero-sum games of linear discrete-time systems via off-policy reinforcement learning, Optim. Control Appl. Methods 41 (4) (2020) 1233–1250.
[36] F.L. Lewis, D. Vrabie, K.G. Vamvoudakis, Reinforcement learning and feedback control: using natural decision methods to design optimal adaptive controllers, IEEE Control Syst. Mag. 32 (6) (2012) 76–105.
[37] L.V. Kantorovich, G.P. Akilov, Funclional analysis in sortned spaces, Mamillan, New York, 1964.
[38] J. Cheng, L. Xie, J.H. Park, et al., Protocol-based output-feedback control for semi-markov jump systems, IEEE Trans. Automat. Contr. 67 (8) (2022a) 4346–4353.
[39] J. Cheng, J.H. Park, H. Yan, et al., An event-triggered round-robin protocol to dynamic output feedback control for nonhomogeneous markov switching systems, Automatica 145 (2022b).
[40] J. Cheng, J.H. Park, Z.G. Wu, A hidden markov model based control for periodic systems subject to singular perturbations, Syst. Control Lett. 157 (2021) 105059.
[41] M. Zhang, C. Shen, Z.G. Wu, Asynchronous observer-based control for exponential stabilization of markov jump systems, IEEE Trans. Circuits Syst. II: Express Briefs 67 (10) (2019b) 2039–2043.
[42] M. Zhang, P. Shi, L. Ma, et al., Quantized feedback control of fuzzy markov jump systems, IEEE Trans. Cybern. 49 (9) (2018) 3375–3384.
[43] M. Zhang, P. Shi, C. Shen, et al., Static output feedback control of switched nonlinear systems with actuator faults, IEEE Trans. Fuzzy Syst. 28 (8) (2019) 1600–1609.

[44] W. Zou, C.K. Ahn, Z. Xiang, Analysis on existence of compact set in neural network control for nonlinear systems, Automatica 120 (2020) 109155.
[45] J. Mao, H.R. Karimi, Z. Xiang, Observer-based adaptive consensus for a class of nonlinear multiagent systems, IEEE Trans. Syst. Man Cybern. 49 (9) (2017) 1893–1900.