Received: 7 February 2021 | Revised: 21 April 2021 | Accepted: 1 June 2021

DOI: 10.1002/rnc.5648

RESEARCH ARTICLE

WILEY

H_∞ optimal control for semi-Markov jump linear systems via TP-free temporal difference (λ) learning

Yaogang Chen | **Jiwei Wen** | **Xiaoli Luan** | **Fei Liu**

Key Laboratory of Advanced Process Control for Light Industry, Institute of Automation, Jiangnan University, Wuxi, China

Correspondence
Xiaoli Luan, Key Laboratory of Advanced Process Control for Light Industry, Institute of Automation, Jiangnan University, Wuxi 214122, China.
Email: xiaoli_luan@126.com

Funding information
General Research Program of Jiangnan University, Grant/Award Number: JUSRP221014; National Natural Science Foundation of China, Grant/Award Numbers: 61722306, 61833007, 61991402, 62073154

Abstract

In the present study, a temporal difference (TD) learning algorithm is proposed to solve the H_∞ optimal control problem for semi-Markov jump linear systems (S-MJLSs). The proposed scheme is TP-free so that it can be applied in cases without pre-known transition probabilities of embedded Markov chain. Coupled algebraic Riccati equations (CAREs) implied with the analytical solution of control gains are derived by utilizing a S-MJLS augmented with maximum sojourn time, which contributes to develop the TD learning algorithm. It is proved that for sufficiently rich enough jumping modes and jumping numbers observed online, the value function in TD algorithm converges to CAREs solutions. Finally, an example is carried out to evaluate the learning capability of TD algorithm and the effectiveness of the proposed control method.

KEYWORDS

H_∞ optimal control, semi-Markov jump systems, sojourn time, temporal difference learning

1 | INTRODUCTION

Stochastic system models are widely applied in diverse engineering and industrial applications, including electric systems,[1,2] chemical industries,[3] and economic systems,[4] to switch the operational modes with respect to abrupt changes. Recently, Markov jump linear systems (MJLSs) have attracted many scholars to model dynamic systems effectively and accurately. In this regard, significant achievements have been obtained on MJLSs in diverse fields, including stability analysis problems,[5-7] filtering problems,[8-10] and control problems.[11-13] Studies show that transition probabilities (TPs) reflect the stochastic transition of systems so that they can be considered as a critical factor. Despite these achievements, it is a challenge to obtain TPs in practical applications. Moreover, in some cases, TPs cannot be given in advance. In order to resolve this problem, researchers tried to control problems of MJLSs through incomplete TPs[14-16] and uncertain TPs.[17-19]

However, considering the memoryless characteristic of Markov chains, TPs of MJLSs have a constant value and change only when the sojourn time of each single subsystem subjects to an exponential distribution (in continuous-time cases) or geometric distribution (in discrete-time cases), indicating the limitation of the MJLS model in practical engineering. To overcome this restriction, semi-Markov jump linear systems (S-MJLSs) have been developed, wherein TPs vary with the historical mode sojourn time. It is worth noting that because of the memory property, S-MJLS is a more general model for simulating industrial systems so that it has been applied in numerous systems, including transportation systems,[20] helicopter systems,[21] and networked mass-spring systems.[22] Recently, required conditions for the system stability have been established.[23-27] Based on the ideal assumption of known transitions, Zhang et al.[28] and Jafari and Ketan[29] developed robust control problems with uncertain system dynamics and optimal control criteria, respectively.

However, the key transitions, called semi-Markov kernel (SMK), consisting of TPs of embedded Markov chain (EMC) and sojourn-time probability density functions (ST-PDFs) are not reachable yet in practical applications. It is worth noting that this shortcoming makes the research of S-MJLSs very difficult and complicated.

In the past few years, the stabilization and control problems have been investigated for S-MJLSs with unavailable transition information. In this regard, stability synthesis with incomplete SMK[30] and unknown TPs or ST-PDFs[31] and dynamic output-feedback control with incomplete SMK[32] have been carried out so far. However, almost all published results have the following two restrictions: (1) Since the transition information is normally discarded, the obtained stability or control results are conservative. (2) Stability conditions are described by linear matrix inequalities (LMIs) so that an explicit controller cannot be acquired. Accordingly, it is an enormous challenge to apply these methods in learning and optimizing methods.

As a branch of reinforcement learning,[33] temporal difference (TD) learning methods such as Q-learning (off-policy TD), Sarsa (on-policy TD), TD(0), and TD(λ), can be applied to combine the advantages of Monte-Carlo and dynamic programming. This intelligent algorithm, which is based on the action-reward principle in the interaction with the environment, is model-free and has been extensively applied to seek the optimal Markov decision process (MDP). Meanwhile, since the optimal control problems can be regarded as MDP, the TD learning method is an appropriate scheme to capture the optimal behavior of system response with no need for prior knowledge about dynamics or stochastic properties of the system. In this regard, the Q-learning algorithm is applied to address the linear quadratic regulator problem[34,35] and H_∞ control problem[36] for discrete-time linear system with unknown dynamic information. In order to solve the nonlinear control problems with uncertain or incomplete system dynamics, some recent new results based on neural network have been reported, such as adaptive optimal control[37-39] and optimal tracking control.[40] Moreover, it is found that an online TD(λ) algorithm can be developed to effectively estimate the solutions of coupled algebraic Riccati equations (CAREs) for the robust control of MJLSs with no need to TPs information.[41]

Inspired by the performed literature survey, it is intended to address H_∞ optimal control problems of S-MJLSs by developing an online TD (λ) algorithm for the case that TPs of EMC are unavailable. The main contributions of the present study can be summarized as follows: First, an augmented S-MJLS is constructed by using the maximum sojourn time as the predictive horizon. Then, the CAREs for H_∞ optimal control of S-MJLSs are presented and explicit controllers are given. Moreover, an improved online TD (λ) learning algorithm is proposed to gather the reward of each mode and optimize the control policy without any prior knowledge about TPs. It is also demonstrated that the value functions finally converge to the solutions of CAREs so that the H_∞ optimal controllers can be obtained. The main advantages of the present article can be summarized as follows:

1. The proposed approach based on the TD (λ) learning does not require apriori TP knowledge of EMC.
2. In the TD (λ) algorithm, the weight vector and the value function are updated at every mode jumping within an episode, thereby increasing the convergence speed.
3. Compared with recent studies on S-MJLSs with incomplete LMI transitions, the proposed scheme is less conservative due to fully online mode observing and is capable of obtaining the explicit controllers based on the CAREs theorem.

The structure of this article is organized as follows: First, definitions and problem descriptions are presented in the following section. Then CAREs conditions are established for the H_∞ optimal control problem of S-MJLSs and then an online TD(λ) algorithm is proposed accordingly. In order to evaluate the performance of the proposed method, simulation results are presented in Section 4, followed by a conclusion and main achievements in Section 5.

Notations: $\mathbb{R}^n$ signifies n-dimensional Euclidean space, $\mathbb{H}^{n+}$ is s set of positive semi-definite n-dimension matrix, A^T and A^{-1} denote the transpose and the inverse of the matrix A, respectively. For $P > 0$ ($P \geq 0$), P is positive (semi-positive) symmetric. $\|A\|$ is the Euclidean norm of matrix A, $F(X)$ is a matrix function in terms of matrix X, $E\{\cdot\}|_x$ stands for the conditional expectation operation conditioned on x, I and O stand for the identity matrix and zero matrix, respectively.

2 | PRELIMINARIES AND PROBLEM FORMULATIONS

We consider a discrete-time stochastic jumping system as follows:

$$x(k+1) = A(r_k)x(k) + B(r_k)u(k) + G(r_k)w(k), \tag{1}$$

where $x(k) \in \mathbb{R}^n$ is the system state, $u(k) \in \mathbb{R}^m$ is the control input, $w(k) \in R^r$ is the disturbance input. Moreover, $\{r_k\}_{k\in\mathbb{N}_+}$ represents a semi-Markov chain (SMC) and takes values in a given stochastic process set $\mathbb{M} = \{1, 2, \dots, N\}$. For $r_k = i \in \mathbb{M}$, A_i, B_i, and G_i are the known matrices, reflecting the system dynamics. The defined system in Equation (1) is called S-MJLSs.

Before giving the definition of SMC, we first review the concepts of Markov renewal chain (MRC),[23] which involves three stochastic processes $\{k_n\}$, $\{R_n\}$, and $\{S_n\}$ are as follows:

1. $\{k_n\}_{n\in\mathbb{N}_+} \in \mathbb{N}_+$, where k_n is the nth jumping time of r_k is a monotonically increasing set of jumping instants, which is a function of the system mode variation.
2. $\{R_n\}_{n\in\mathbb{N}_+} \in \mathbb{M}$, where R_n is the mode index of r_k at the nth jump denotes a mode sequence.
3. $\{S_n\}_{n\in\mathbb{N}_+} \in \mathbb{N}_+$, where $S_n = k_n - k_{n-1}$ is the mode sojourn time between consequent adjacent jumps presents the sequence of mode sojourn time.

Definition 1. $\{(R_n, k_n)\}_{n\in\mathbb{N}_+}$ is a discrete-time MRC if $\forall i, j \in \mathbb{M}$, $\tau \in \mathbb{N}_+$, $n \in \mathbb{N}_+$.[23] Under this circumstance, the TP of mode jumps is as follows:

$$\Pr(R_{n+1} = j, S_{n+1} = \tau | R_0, \dots, R_n = i; k_0, \dots, k_n) = \Pr(R_{n+1} = j, S_{n+1} = \tau | R_n). \tag{2}$$

In this case, $\{R_n\}_{n\in\mathbb{N}_+}$ can be considered as an EMC. Moreover, the transition characteristics of MRC and EMC can be described through the following probability functions:

1. The TP matrix of $\{R_n\}_{n\in\mathbb{N}_+}$ is defined as $\Theta = [\theta_{ij}]_{i,j\in\mathbb{M}}$, where $\theta_{ij} = \Pr(R_{n+1} = j | R_n = i)$ and $\theta_{ii} = 0$.
2. The discrete-time SMK is defined as $\Pi = [\pi_{ij}]_{i,j\in\mathbb{M}}$, where $\pi_{ij}(\tau) = \Pr(R_{n+1} = j, S_{n+1} = \tau | R_n = i)$ and $\sum_{\tau=0}^{\infty}\sum_{j\in\mathbb{M}} \pi_{ij}(\tau) = 1$ with $\pi_{ij}(0) = 0$.
3. The ST-PDF of the i-mode is defined as $w_{ij}(\tau) = \Pr(S_{n+1} = \tau | R_n = i, R_{n+1} = j)$, $\forall i, j \in \mathbb{M}$, $\forall \tau \in \mathbb{N}_+$.

Then the relationship between $\pi_{ij}(\tau)$, $w_{ij}(\tau)$ and θ_{ij} can be expressed in the form below:

$$\pi_{ij}(\tau) = w_{ij}(\tau)\theta_{ij}. \tag{3}$$

Definition 2. Based on the definition of MRC, $\{r_k\}_{k\in\mathbb{N}_+}$ can be considered as SMC, when the following condition holds:[23]

$$r_k = R_n, \forall k \in [k_n, k_{n+1}) \in \mathbb{N}_+, n \in \mathbb{N}_+. \tag{4}$$

Definition 3. For an initial state $x_0 \in \mathbb{R}^n$, initial mode $r_0 \in \mathbb{M}$ and upper bound $T^i_{\max} \in \mathbb{N}_+$ of the sojourn time for $i \in \mathbb{M}$ and $w \equiv 0$, the defined system in the form of Equation (1) is σ-error mean square stable (σ-MSS) if the following condition holds[23]: $u \equiv 0$

$$\lim_{k\to\infty} E\left[\|x(k)\|^2\right]\Big|_{x_0, r_0, \tau_i \le T^i_{\max}} = 0, \tag{5}$$

with

$$\sigma \triangleq \sum_{i\in\mathbb{M}} \left| \ln\left[\sum_{\tau=0}^{T^i_{\max}} \sum_{j\in\mathbb{M}} \pi_{ij}(\tau)\right]\right|, \tag{6}$$

where τ_i is the sojourn time of the i-mode and σ denotes the approximation error of σ-MSS to real MSS.

Then the optimal performance index with H_∞ attenuation level γ is considered in the form below:

$$J^*(x_k) = \min_u \max_w E\left[\sum_{k=k_n}^{\infty} \left(x_k^T Q_i x_k + u_k^T u_k - \gamma^2 w_k^T w_k \,|r_{k_n} = i, i \in \mathbb{M}\right)\right] \tag{7}$$

for a fixed value γ in the H_∞ attenuation level of system (1) with $Q_i > 0, \forall i \in \mathbb{M}$.

The main objective of the present article is to design a control scheme for system (1) so that the closed-loop system becomes σ-MSS with the optimal performance index defined in (7).

Remark 1. Generally, $T^i_{\max}$ could be infinite, which originates from its high calculation and derivation complexities. In order to resolve this problem, it is assumed that $T^i_{\max}$ is a finite constant. Although this assumption inevitably leads to errors defined in (6), it indicates that as $T^i_{\max}$ increases, the corresponding σ decreases $\forall i \in \mathbb{M}$. This can also be mathematically expressed as $T^i_{\max} \to \infty$, then $\sigma \to 0$. It is inferred that finite $T^i_{\max}$ has a negligible impact on actual system performance if $\eta_i = \sum_{\tau=0}^{T^i_{\max}} \sum_{j\in\mathbb{M}} \pi_{ij}(\tau)$ approaches probability 1.

Remark 2. Formula (3) indicates that unlike MJLS with a constant TP matrix, SMK of S-MJLS is not only related to TPs of EMC but also it depends on the ST-PDFs. In other words, it is necessary to consider an extra time dimension to obtain the transitions of S-MJLS with the following limitations:

1. One-step TP of $\{r_k\}$, which is defined as $\epsilon_{ij}(\vec{k}) \triangleq \Pr(r_{k+1} = j \,|\, r_k = i)$, $i,j \in \mathbb{M}$ is a history dependent function, and requires an infinite number of computations.[24] Accordingly, it is rather a challenge to directly use ϵ_{ij} of system (1) and develop a similar desired control scheme.
2. In many practical applications, transitions between different system modes are unknown. Accordingly, the SMK cannot be calculated in advance. Consequently, it is a challenge to obtain feasible solutions of existing LMI conditions.

These problems will be further addressed in the following sections.

3 | MAIN RESULTS

In this section, it is intended to initially establish the H_∞ optimal control conditions described by CAREs, based on the augmented S-MJLSs. Then a TD-learning algorithm is designed to approximate the TP-related solutions of the presented CAREs without TPs of EMC. Finally, the convergence rate of the proposed algorithm is studied in an online case study.

3.1 | CARE conditions for H_∞ optimal control problems of S-MJLSs

In this part, the explicit control law is developed for S-MJLS H_∞ optimal control problems. To overcome the limitations mentioned in Remark 2, the following augmented S-MJLSs are present based on the maximum sojourn time $T^i_{\max}$, $\forall i \in \mathbb{M}$:

$$\overline{X}_{i,k_n} = \bar{A}_i x_{k_n} + \overline{B}_i U_{i,k_n} + \overline{G}_i W_{i,k_n}, \tag{8}$$

where $\overline{X}_{i,k_n} = \begin{pmatrix} x_{k_n+1} \\ x_{k_n+2} \\ \vdots \\ x_{k_n+T^i_{\max}} \end{pmatrix}$, $U_{i,k_n} = \begin{pmatrix} u_{k_n} \\ u_{k_n+1} \\ \vdots \\ u_{k_n+T^i_{\max}-1} \end{pmatrix}$, $W_{i,k_n} = \begin{pmatrix} w_{k_n} \\ w_{k_n+1} \\ \vdots \\ w_{k_n+T^i_{\max}-1} \end{pmatrix}$, $\bar{A}_i = \begin{pmatrix} A_i \\ A_i^2 \\ \vdots \\ A_i^{T^i_{\max}} \end{pmatrix}$, $\overline{B}_i = \begin{pmatrix} B_i & & & \\ A_iB_i & B_i & & \\ \vdots & & \ddots & \\ A_i^{T^i_{\max}-1}B_i & \dots & & B_i \end{pmatrix}$,

$\overline{G}_i = \begin{pmatrix} G_i & & & \\ A_iG_i & G_i & & \\ \vdots & & \ddots & \\ A_i^{T^i_{\max}-1}G_i & \dots & & G_i \end{pmatrix}$.

Then the CAREs conditions can be presented for H_∞ optimal control problems based on the augmented S-MJLSs.

Theorem 1. *Given a scalar $\gamma > 0$ and the maximum sojourn time $T^i_{\max} \in \mathbb{N}_+$ for all $i \in \mathbb{M}$, there exists a control law $U_{i,k_n} = \overline{F}_i x_{k_n}$ so that the system (1) is σ-MSS and satisfies an H_∞ optimal performance index (7), if there exists $P_i \in \mathbb{H}^{n+}$ satisfying the following CAREs:*

$$P_i = \left(\bar{A}_i + \overline{B}_i\overline{F}_i + \overline{G}_i\overline{K}_i\right)^T \overline{P}_i \left(\bar{A}_i + \overline{B}_i\overline{F}_i + \overline{G}_i\overline{K}_i\right) + \overline{F}_i^T \bar{I}_i \overline{F} - \gamma^2 \overline{K}_i^T \bar{I}_i \overline{K}_i, \tag{9}$$

$$\gamma^2 I - \overline{G}_i^T \overline{P}_i \overline{G}_i > 0, \bar{I}_i + \overline{B}_i^T \overline{P}_i \overline{B}_i > 0, \tag{10}$$

where

$$\overline{P}_i = \frac{1}{\eta_i}\begin{pmatrix} (\eta_i - \pi_i(1))\, Q_i + \sum_{j\in\mathbb{M}} \pi_{ij}(1)\left(P_j + Q_j\right) & & & \\ & (\eta_i - \pi_i(1) - \pi_i(2))\, Q_i + \sum_{j\in\mathbb{M}} \pi_{ij}(2)\left(P_j + Q_j\right) & & \\ & & \ddots & \\ & & & \left(\eta_i - \sum_{\tau=1}^{T^i_{\max}-1} \pi_i(\tau)\right) Q_i + \sum_{j\in\mathbb{M}} \pi_{ij}(T^i_{\max}-1)\left(P_j + Q_j\right) \\ & & & \sum_{j\in\mathbb{M}} \pi_{ij}(T^i_{\max})\left(P_j + Q_j\right) \end{pmatrix} \tag{11}$$

and

$$\bar{I}_i = \frac{1}{\eta_i}\begin{pmatrix} \eta_i I & & & \\ & (\eta_i - \pi_i(1))\, I & & \\ & & \ddots & \\ & & & \left(\eta_i - \sum_{\tau=1}^{T^i_{\max}-1} \pi_i(\tau)\right) I \end{pmatrix}, \tag{12}$$

where $\pi_i(\tau) = \sum_{j\in\mathbb{M}} \pi_{ij}(\tau)$. *Then, the* H_∞ *optimal augmented controller* $\overline{F}_i$ *and the disturbance input gain* $\overline{K}_i$ *for* $i \in \mathbb{M}$ *can be calculated through the following expressions:*

$$\begin{aligned} \overline{F}_i(\overline{P}_i) = &-\left(\bar{I}_i + \overline{B}_i^T \overline{P}_i \overline{B}_i + \overline{B}_i^T \overline{P}_i \overline{G}_i \left(\gamma^2 \bar{I}_i - \overline{G}_i^T \overline{P}_i \overline{G}_i\right)^{-1} \overline{G}_i^T \overline{P}_i \overline{B}_i\right)^{-1} \\ &\times \left(\overline{B}_i^T \overline{P}_i \bar{A}_i + \overline{B}_i^T \overline{P}_i \overline{G}_i \left(\gamma^2 \bar{I}_i - \overline{G}_i^T \overline{P}_i \overline{G}_i\right)^{-1} \overline{G}_i^T \overline{P}_i \bar{A}_i\right), \end{aligned} \tag{13}$$

$$\begin{aligned} \overline{K}_i(\overline{P}_i) = &\left(\gamma^2 \bar{I}_i - \overline{G}_i^T \overline{P}_i \overline{G}_i + \overline{G}_i^T \overline{P}_i \overline{B}_i \left(\bar{I}_i + \overline{B}_i^T \overline{P}_i \overline{B}_i\right)^{-1} \overline{B}_i^T \overline{P}_i \overline{G}_i\right)^{-1} \\ &\times \left(\overline{G}_i^T \overline{P}_i \bar{A}_i - \overline{G}_i^T \overline{P}_i \overline{B}_i \left(\bar{I}_i + \overline{B}_i^T \overline{P}_i \overline{B}_i\right)^{-1} \overline{B}_i^T \overline{P}_i \bar{A}_i\right). \end{aligned} \tag{14}$$

Proof. The infinite-horizon of H_∞ optimal performance index (7) can be written as follows:

$$J^*(x_{k_n}) = \min_u \max_w E\left[\sum_{k=k_n}^{k_{n+1}-1} \left(x_k^T Q_i x_k + u_k^T u_k - \gamma^2 w_k^T w_k\right) + J^*(x_{k_{n+1}}) | r_{k_n} = i, i \in \mathbb{M}\right], \tag{15}$$

which can be demonstrated to be equivalent to the following expression:

$$V^*\left(x_{k_n}, u_{k_n}, w_{k_n}\right) = \min_u \max_w V\left(x_{k_n}, u_{k_n}, w_{k_n}\right), \tag{16}$$

$$\begin{aligned} V\left(x_{k_n}, u_{k_n}, w_{k_n}\right) = E\Bigg[&\sum_{k=k_n+1}^{k_{n+1}} \left(x_k^T Q_i x_k\right) + \sum_{k=k_n}^{k_{n+1}-1} \left(u_k^T u_k - \gamma^2 w_k^T w_k\right) \\ &+ V\left(x_{k_{n+1}}, u_{k_{n+1}}, w_{k_{n+1}}\right) | r_{k_n} = i, i \in \mathbb{M}\Bigg]. \end{aligned} \tag{17}$$

Defining $V^*\left(x_{k_n}, u_{k_n}, w_{k_n}\right) = x_{k_n}^T P_i x_{k_n}$, using the augmented systems (8), $U_{i,k_n} = \overline{F}_i x_{k_n}$ and $W_{i,k_n} = \overline{K}_i x_{k_n}$ in (17), the result of CAREs (9) can be obtained. It is worth noting that based on optimality conditions $\frac{\partial V}{\partial U_{i,k_n}} = 0$, $\frac{\partial V}{\partial W_{i,k_n}} = 0$, $\frac{\partial^2 V}{\partial U_{i,k_n}^2} > 0$, $\frac{\partial^2 V}{\partial W_{i,k_n}^2} < 0$, inequality (10) holds. Meanwhile, an analytical expression can be obtained for the desired controller (13) and the maximum disturbance gain (14), thereby completing the proof. ∎

Remark 3. The augmented S-MJLS (8), which contains all possible sojourn times, sojourn time-dependent control signals and disturbances can reveal the correlation between the S-MJLS dynamics and the sojourn time in the i th mode. In this

case, $\overline{F}_i = \left(F_{i,1}, F_{i,2}, \ldots, F_{i,T^i_{\max}}\right)$, where $F_{i,\tau}$, $\tau \in [1, T^i_{\max}]$ is the sojourn time-dependent feedback controller defined by $u_{k_n+\tau-1} = F_{i,\tau}x_{k_n}$. It is worth noting that the feedback mechanism of the augmented S-MJLS is based on the states of jump moments x_{k_n} rather than the states of time steps x_k.

Remark 4. The main advantage of CAREs compared to conventional methods for transforming system stability conditions into LMIs is obtaining explicit controllers for policy iterative learning. This feature can be applied in the sequel parts.

3.2 | Online TD learning algorithm for S-MJLS

In this section, the proposed CAREs formulation is applied to develop the TD-learning algorithm for S-MJLS. The main advantage of the TD-learning over conventional control methods for solving CAREs and LMIs is that the TPs θ_{ij} of EMC in the TD-learning can be unknown, no matter it is a fixed or a time-varying parameter.

Assumption 1. For simplicity, it is assumed that $Q_i = Q, \forall i \in \mathbb{M}$, and the PDF of the sojourn time from i th mode to other modes satisfies the same distribution, that is

$$w_{ij}(\tau) = w_i(\tau), \forall i, j \in \mathbb{M}, i \neq j, \forall \tau \in \mathbb{N}_+. \tag{18}$$

Then, the mode-dependent value function of TD learning can be updated in the form below:

$$Y_i(t, n+1) = Y_i(t, k) + \gamma_i(t)e_i(t, n)d(t, n), \tag{19}$$

$$Y_i(t, 0) = Y_i(t), Y_i(t+1) = Y_i(t, N(t)), \tag{20}$$

where n is the number of mode jumps of EMC and t (episode) is the sequence number of the mode trajectories. Moreover, $N(t)$ denotes the length of mode trajectory of EMC in the tth episode. The stepsize $\gamma_i(t)$ and eligibility trace $e_i(t, n)$ satisfy the following conditions:

$$\sum_{t=0}^{\infty} \gamma_i(t) = \infty, \sum_{t=0}^{\infty} \gamma_i^2(t) < \infty, \tag{21}$$

$$e_i(t, n) = \begin{cases} 0, & n < n_i(t), \\ \lambda^{n-n_i(t)}, & n \geq n_i(t), \end{cases} \tag{22}$$

where $n_i(t) = \inf_n\{r(t, n) = i\}$, $i \in \mathbb{M}$ and $0 < \lambda < 1$. Based on Assumption 1, we have $\pi_i(\tau) = w_i(\tau), \forall i \in \mathbb{M}$. Defining reward functions $N_i = \overline{F}_i^{\mathrm{T}}\bar{I}_i\overline{F}_i - \gamma^2\overline{K}_i^{\mathrm{T}}\bar{I}_i\overline{K}_i$ and assuming $S_i = \bar{A}_i + \overline{B}_i\overline{F}_i + \overline{G}_i\overline{K}_i$, the TD error $d(t, k)$ can be calculated as:

$$d(t, n) = N_{r(t,n+1)} + \frac{1}{\eta_{r(t,n+1)}} S^{\mathrm{T}}_{r(t,n+1)}\Lambda(Y_{r(t,n)}(t, n))S_{r(t,n+1)} - Y_{r(t,n)}(t, n), \tag{23}$$

where

$$\Lambda(Y_{r(t,n)}(t, n)) = \begin{pmatrix} w_{r(t,n+1)}(1)Y_{r(t,n+1)}(t, n) & & \\ & w_{r(t,n+1)}(2)Y_{r(t,n+1)}(t, n) & \\ & & \ddots \\ & & w_{r(t,n+1)}(T^{r(t,n+1)}_{\max})Y_{r(t,n+1)}(t, n) \end{pmatrix} + \begin{pmatrix} \eta_{r(t,n+1)}Q & & \\ & \left(\eta_{r(t,n+1)} - w_{r(t,n+1)}(1)\right)Q & \\ & & \ddots \\ & & \left(\eta_{r(t,n+1)} - \sum_{\tau=1}^{T^{r(t,n+1)}_{\max}-1} w_{r(t,n+1)}(\tau)\right)Q \end{pmatrix}. \tag{24}$$

Theorem 2. *Suppose that the initial control policies $\overline{F}_i^0$ and $\overline{K}_i^0$ stabilize the system (8) and Assumption 1 holds, there exists a solution $P_i > 0$ for CARE (9). Then, apply Algorithm 1, we have $\lim_{t\to\infty} Y_i(t) = \sum_{j\in\mathbb{M}} \theta_{ij}P_j$ and $\lim_{t\to\infty} \overline{F}_i(\Lambda(Y_i(t))) = \overline{F}_i(\overline{P}_i)$, which makes system (1) σ-MSS with an H_∞ optimal performance.*

Algorithm 1. TD learning for S-MJLS with no TPs information of EMC

Initialization

Start with initial control policies $\overline{F}_i^0$ and $\overline{K}_i^0$, which make the system (8) stable,
$Y_{i(0)}^0 = \mathbf{0}, \forall i \in \mathbb{M}$, and the iteration number is $l = 0$.

Policy evaluation

1: **for** $t = 1, \ldots, T$ **do**
2: Initialize the random initial mode $r(t,0) = i$ and $e_i(t,0) = 0$,
3: Execute $\overline{F}_i = F_i^l, \overline{K}_i = K_i^l, \forall i \in \mathbb{M}$,
4: **for** $n = 1, \ldots, N(t)$ **do**
5: Observe the online SMC $\{R_{n+1}, k_{n+1}\}$, recode the next mode $r(t, n+1)$,
6: Update $e_i(t,n)$ by (22),
7: Update $Y_i(t, n+1)$ by (19),
8: **end for**
9: **end for**
10: $Y_i^{l+1} = Y_i(T+1)$,

Policy improvement

11: $\Lambda_i^{l+1} = \Lambda(Y_i^{l+1})$ via (24),
12: $\overline{F}_i^{l+1} = \overline{F}_i(\Lambda_i^{l+1})$ via (13),
13: $\overline{K}_i^{l+1} = \overline{K}_i(\Lambda_i^{l+1})$ via (14),

Unless $\left\|Y_i^l - Y_i^{l-1}\right\| < \epsilon, \forall i \in \mathbb{M}$ for a small positive value of ϵ, set $l = l+1$ and go to step 1.

Proof. Based on Lemma 2 and the proof of Theorem 1 in Reference 41, the proof that $Y_i(t)$ converges to $\sum_{j\in\mathbb{M}} \theta_{ij} P_j, \forall i \in \mathbb{M}$ when $t \to \infty$ consists of two parts: First, an offline TD value function $\tilde{Y}_i(t)$ is defined, which is updated every time after observing a complete mode trajectory and converges to $\sum_{j\in\mathbb{M}} \theta_{ij} P_j$. Then, it should be proved that the online form $Y_i(t)$ and the offline form $\tilde{Y}_i(t)$ converge to the same value. In this regard, $\tilde{Y}_i(t)$ is defined as follows:

$$\tilde{Y}_i(t+1) = \tilde{Y}_i(t) + \sum_{n=0}^{N(t)-1} \gamma_i(t) e_i(t,n) \tilde{d}(t,n), \tag{25}$$

$$\tilde{d}(t,n) = N_{r(t,n+1)} + \frac{1}{\eta_{r(t,n+1)}} S_{r(t,n+1)}^{\mathrm{T}} \Lambda(\tilde{Y}(t)) S_{r(t,n+1)} - \tilde{Y}_{r(t,n)}(t). \tag{26}$$

It can be proved that the conditions (a), (b), and (c) of Lemma 2 in Reference 41 are satisfied, which completes the first step of the proof. Then the offline TD value function $\tilde{Y}_i(t)$ can be written as the following one-step form:

$$\tilde{Y}_i(t, n+1) = \tilde{Y}_i(t,k) + \gamma_i(t) e_i(t,n) \tilde{d}(t,n), \tag{27}$$

$$\tilde{Y}_i(t+1) = \tilde{Y}_i(t, N(t)). \tag{28}$$

Introducing (22) into (27) results in the following expression:

$$\left\|\tilde{Y}_i(t,n) - \tilde{Y}_i(t,0)\right\| \le \xi n \gamma_i(t), \tag{29}$$

where ξ is a bounded constant. Based on the mathematical induction, we obtain the following inequality:

$$\left\|Y_i(t,n) - \tilde{Y}_i(t,n)\right\| \le V(n) \gamma_{\max}^2(t), \tag{30}$$

where $V(n+1) = V(n)(1 + 2\gamma_{\max}(t)) + 2\xi N(t)$ and $\gamma_{\max}(t) = \max_i \gamma_i(t)$. Combining (27)–(30) yields the following expression:

$$\left\| \frac{\tilde{Y}_i(t+1) - Y_i(t+1)}{\gamma_{\max}^2(t+1)} \right\| \le V(N(t+1)). \tag{31}$$

When $t \to \infty$, (21) indicates that $\gamma_{\max}^2(t) = 0$. Meanwhile, $N(t)$ is a finite function so that $V(N(t))$ is bounded, then $\left\|\tilde{Y}_i(t) - Y_i(t)\right\| = 0$ and the proof is completed. ∎

TABLE 1 Economic transients with α and s

Mode i	Economic transient	Coefficients α, s
1	Norm	$\alpha = 2.5, s = 0.3$
2	Boom	$\alpha = 4.3, s = 0.8$
3	Slump	$\alpha = -5.3, s = 0.9$

Remark 5. Algorithm 1 is developed by the policy iteration framework, including the initialization, policy evaluation (PE), and policy improvement (PI) steps. More specifically, the algorithm begins with an initial value function Y_i^0 and feasible control policies, which stabilize the closed-loop system. Then in the PE step $Y_i(t,n)$ is updated by online observations of modes and its jump numbers of MRC until it converges. Finally, a new control policy is calculated in the PI step via the convergent Y_i^{l+1}. In this algorithm, Λ_i^l and F_i^l approximate $\overline{P}_i$ and the desired optimal controller, respectively.

Remark 6. The trace-decay parameter λ determines the convergence rate, at which the eligibility trace falls. It is worth noting that the smaller the λ parameter, the higher the algorithm convergence accuracy, but the lower the convergence speed. The final convergence value of the algorithm depends on the TD error $d(t,n)$, which is designed by CARE (9).

Remark 7. In the present study, the case that TPs of EMC are unavailable while the ST-PDFs are available is addressed. Unlike conventional methods, the proposed approach does not solve CAREs (9), but obtains the solutions by online iteration learning and approximating. Consequently, there is no need for the TP information between system modes.

Remark 8. The proposed algorithm has some differences from the previous TD algorithm for MJLS.[41] These differences are as the following:

1. In the proposed algorithm, TPs of EMC are unknown, which is different from the case with unavailable one-step TPs. Therefore, the minimum updating unit n denotes the mode jumping number of EMC rather the than time step of the system.
2. Since the one-step TP under S-MJLS is a function of the modes sojourn time, it is necessary to consider all possible cases of sojourn time of each system mode in Algorithm 1. It should be indicated that the obtained control policy in the proposed algorithm is a sequence consisting of all possible sojourn time-dependent controllers in the i th mode.

4 | ILLUSTRATIVE EXAMPLE

In this part, it is intended to evaluate the effectiveness of the proposed approach. In this regard, a comparative simulation is carried out between the proposed CAREs method (with known TPs of EMC) and the TD-learning approach (with unknown TPs of EMC). The Samuelson's macroeconomic model[42] is utilized, which takes advantage of multiplier analysis and accelerator principle to predict changes in economic dynamics. Studies[43] reveal that the Samuelson's model can be expressed in a state space with disturbances:

$$x(k+1) = \begin{bmatrix} 0 & 1 \\ -\alpha & 1-s+\alpha \end{bmatrix} x(k) + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u(k) + \begin{bmatrix} 0 \\ 0.03 \end{bmatrix} w(k), \tag{32}$$

where $x(k)$, $u(k)$, and $w(k)$ denote the national income, government expenditure, and net export capital, respectively. Based on the issued financial data by the U.S. Department of Commerce, the accelerator coefficient α and the multiplier coefficient s can be calculated in three groups. Accordingly, systems can be divided into three modes, including norm, boom, and slump. Table 1 presents the correlation between α, s and economic transients.

Further investigations[44] reveal that the economic fluctuation regularity of the time cycle of booms and recessions is quite intricate so that switching probabilities of economic transients may change over time. Accordingly, it is inferred that S-MJLSs are more appropriate schemes for modeling these systems when the comparison is made with MJLSs. Based on Assumption 1, ST-PDFs of all modes can be obtained by $\omega_1(\tau) = 0.4^{\tau-1^{1.3}} - 0.4^{\tau^{1.3}}$, $\omega_2(\tau) = 0.3^{\tau-1^{0.8}} - 0.3^{\tau^{0.8}}$,

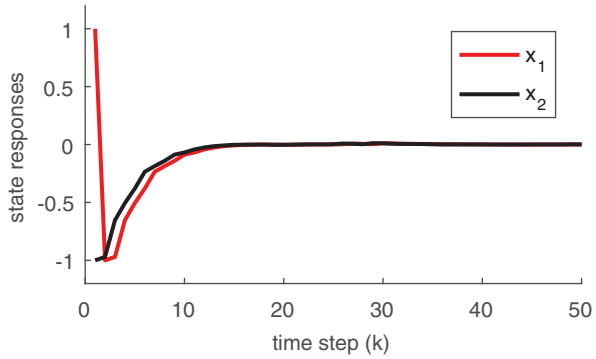

FIGURE 1 Evolutions of the system state $x(k)$ subjected to the CARE approach with known TPs [Colour figure can be viewed at wileyonlinelibrary.com]

$\omega_3(\tau) = 0.5^{\tau-1}e^{-0.5}/(\tau-1)!$. Meanwhile, the TP matrix of EMC can be expressed in the form below:

$$\Theta = \begin{bmatrix} 0 & 0.5072 & 0.4928 \\ 0.5357 & 0 & 0.4643 \\ 0.1507 & 0.8493 & 0 \end{bmatrix}. \tag{33}$$

The maximum sojourn time for the system modes is set to $T^1_{max} = 3$, $T^2_{max} = 4$, $T^3_{max} = 3$. Moreover, it can be proved that $\eta_1 = 0.9781$, $\eta_2 = 0.9740$, $\eta_i = 0.9856$, indicating that these probabilities are close to unity. Accordingly, small σ-error is ensured. Considering pre-known TP information of EMC, weight matrix $Q_i = I$ and H_∞ performance index $\gamma = 1$, CAREs (9) can be solved through a recursive algorithm. Accordingly, the sojourn time-dependent H_∞ optimal controllers $\overline{F}_i^{CARE} = \left(F_{i,1}^{CARE}, F_{i,2}^{CARE}, \ldots, F_{i,T^i_{max}}^{CARE}\right)^T$ can be defined as the following: $\overline{F}_1^{CARE} = \begin{pmatrix} 2.4087 & -2.4248 \\ 0.2408 & 0.5320 \\ -0.0517 & 0.6168 \end{pmatrix}$, $\overline{F}_2^{CARE} = \begin{pmatrix} 4.1470 & -3.6794 \\ 0.6135 & 1.1049 \\ -0.1649 & 1.4108 \\ -0.2608 & 1.1500 \end{pmatrix}$, $\overline{F}_3^{CARE} = \begin{pmatrix} -5.1351 & 5.7465 \\ 1.0472 & -2.1021 \\ -0.4184 & 0.0057 \end{pmatrix}$. For a given initial state $x(0) = \begin{bmatrix} 1 & -1 \end{bmatrix}^{\mathrm{T}}$ and disturbance $w(k) = 0.1e^{-0.08k}\sin(0.01k\pi + 0.1\pi)$, the closed-loop response of the Samuelson's macroeconomic system under the obtained controller gains with 50 generated jumping steps are shown in Figure 1. It is observed that both curves converge to zero before $k = 20$ so that σ-MSS of the closed-loop system is ensured. Accordingly, effectiveness of controllers by the CAREs approach is verified.

Then, it is intended to evaluate the validity and accuracy of the TP-free TD(λ) algorithm. The main advantage of this approach is that it is no longer necessary to take TP information as a prior known condition. In this regard, Algorithm 1 is applied in a case study where $T = 400$, $N(t) = 10$, $\lambda = 0.1$, $\gamma_i(t) = 1/t, \forall i \in \mathbb{M}$. Figure 2 illustrates the obtained approximation results from the value function in three modes with 400 trajectories. It is observed that all three value-functions converge quickly. This may be attributed to the online iteration mechanism of the TD(λ) algorithm, which is updated in each mode jump step of the EMC. Accordingly, the output controller matrix can be obtained as $\overline{F}_1^{TD} = \begin{pmatrix} 2.4022 & -2.4162 \\ 0.2400 & 0.5270 \\ -0.0513 & 0.6156 \end{pmatrix}$, $\overline{F}_2^{TD} = \begin{pmatrix} 4.1391 & -3.6646 \\ 0.6019 & 1.0704 \\ -0.1704 & 1.4819 \\ -0.2443 & 1.0721 \end{pmatrix}$, $\overline{F}_3^{TD} = \begin{pmatrix} -5.1206 & 5.7321 \\ 1.0486 & -2.0881 \\ -0.4170 & -0.0152 \end{pmatrix}$. In order to verify the accuracy of the convergence value of the proposed TD(λ) algorithm, controller error is defined in the form below:

$$\Delta_i = \sum_{p,q} \frac{\left\| \left[\overline{F}_i^{TD}\right]_{pq} - \left[\overline{F}_i^{CARE}\right]_{pq} \right\|}{\left\| \left[\overline{F}_i^{CARE}\right]_{pq} \right\|}, \tag{34}$$

where $\left[\overline{F}_i\right]_{pq}$ denotes an element in row p and column q of $\overline{F}_i$. Based on the performed calculations, $\Delta_1 = 0.0033$, $\Delta_2 = 0.0196$, $\Delta_3 = 0.0030$ indicating powerful and precise learning capabilities of Algorithm 1. Applying these controllers for the same initial condition and disturbance, the state trajectories of the closed-loop system (32) can be obtained. Figure 3 shows that the state evolutions of $x(k)$ are similar to the curses in Figure 2, thereby verifying the validity and applicability of the proposed TD(λ) learning approach.

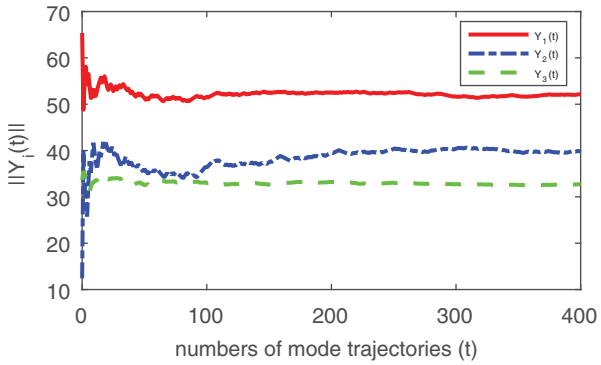

FIGURE 2 Approximating curve of $\|Y_i(t)\|$ in three modes [Colour figure can be viewed at wileyonlinelibrary.com]

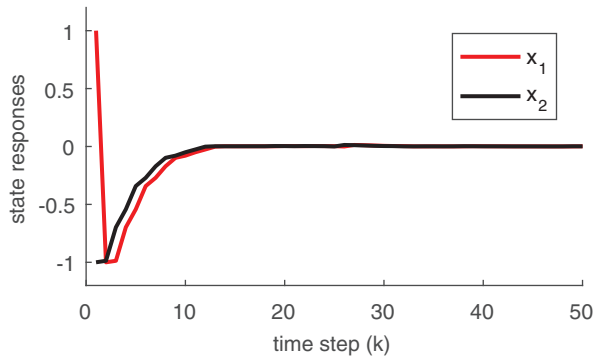

FIGURE 3 Evolution of the system state $x(k)$ subjected to TD(λ) approach with unknown TPs [Colour figure can be viewed at wileyonlinelibrary.com]

5 | CONCLUSION

In the present study, a TP-free TD learning approach is developed to address the H_∞ optimal control problem for discrete-time S-MJLSs. In this regard, CAREs conditions are initially developed through an augmented S-MJLS, which includes the analytical solution of optimal controllers. Then, the TD(λ) algorithm is designed to approximate the control policy quickly, and obtain the desired optimal controller from the jumping modes. Accordingly, the proposed scheme can be entirely separated from TPs information of the EMC. Finally, the proposed method is applied in a microeconomic system to evaluate its effectiveness. It should be indicated that the ST-PDFs are assumed known in all calculations. Accordingly, possible future research is to investigate the real-time sojourn time mode and develop a modified learning algorithm with incomplete ST-PDFs.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grant Nos. 61722306, 61833007, 61991402, and 62073154) and General Research Program of JIANGNAN University (No. JUSRP221014).

CONFLICT OF INTEREST

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

DATA AVAILABILITY STATEMENT

The data availability statement is as follows: (1) The system model parameters in the illustrative example part are obtained from References 34 and 36. (2) The system mode data used to the TD learning algorithm is generated by MATLAB program. No other data is available.

ORCID

Jiwei Wen https://orcid.org/0000-0001-8780-4762
Xiaoli Luan https://orcid.org/0000-0002-4805-1726

REFERENCES

1. Vargas AN, Costa EF, Acho L, do Val JBR. Switching stochastic nonlinear systems with application to an automotive throttle. *IEEE Trans Autom Control*. 2018;63(9):3098-3104.
2. Oliveira RCLF, Vargas AN, Val JBR, Peres PLD. Mode-independent $\mathcal{H}_2$-control of a DC motor modeled as a Markov jump linear system. *IEEE Trans Control Syst Technol*. 2014;22(5):1915-1919.
3. Goh L, Chen K, Bhamidi V, et al. A stochastic model for nucleation kinetics determination in droplet-based microfluidic systems. *Cryst Growth Des*. 2010;10(6):2515-2521.
4. Baten A, Khalid R. Extended optimal stochastic production control model with application to economics. *J Intell Fuzzy Syst*. 2017;32(3):1847-1854.
5. Bolzern P, Colaneri P, De Nicolao G. On almost sure stability of continuous-time Markov jump linear systems. *Automatica*. 2006;42(6):983-988.
6. Hou T, Ma H. Exponential stability for discrete-time infinite Markov jump systems. *IEEE Trans Autom Control*. 2015;61(12):4241-4246.
7. Luan X, Huang B, Liu F. Higher order moment stability region for Markov jump systems based on cumulant generating function. *Automatica*. 2018;93:389-396.
8. Wan H, Luan X, Karimi HR, Liu F. High-order moment filtering for Markov jump systems in finite frequency domain. *IEEE Trans Circuits Syst II Exp Briefs*. 2018;66(7):1217-1221.
9. Zhang L, Boukas E-K. Mode-dependent H_∞ filtering for discrete-time Markovian jump linear systems with partly unknown transition probabilities. *Automatica*. 2009;45(6):1462-1467.
10. Wu Z, Su H, Chu J. H∞ filtering for singular Markovian jump systems with time delay. *Int J Robust Nonlinear Control*. 2010;20(8):939-957.
11. Xiao N, Xie L, Fu M. Stabilization of Markov jump linear systems using quantized state feedback. *Automatica*. 2010;46(10):1696-1702.
12. Wan H, Luan X, Karimi H, Liu F. Dynamic self-triggered controller co-design for Markov jump systems. *IEEE Trans Autom Control*. 2020.
13. Shi P, Liu M, Zhang L. Fault-tolerant sliding-mode-observer synthesis of Markovian jump systems using quantized measurements. *IEEE Trans Ind Electron*. 2015;62(9):5910-5918.
14. Zhang L, Boukas E-K. H∞ control for discrete-time Markovian jump linear systems with partly unknown transition probabilities. *Int J Robust Nonlinear Control IFAC-Affiliat J*. 2009;19(8):868-883.
15. Zong G, Yang D, Hou L, Wang Q. Robust finite-time H_∞ control for Markovian jump systems with partially known transition probabilities. *J Frankl Inst*. 2013;350(6):1562-1578.
16. Shen M, Shen Y, Tang Z, Zhou G. Finite-time H_∞ filtering of Markov jump systems with incomplete transition probabilities: a probability approach. *IET Signal Process*. 2015;9(7):572-578.
17. Luan X, Zhao S, Liu F. H∞ control for discrete-time Markov jump systems with uncertain transition probabilities. *IEEE Trans Autom Control*. 2012;58(6):1566-1572.
18. Faraji-Niri M, Jahed-Motlagh M-R, Barkhordari-Yazdi M. Stochastic stability and stabilization of a class of piecewise-homogeneous Markov jump linear systems with mixed uncertainties. *Int J Robust Nonlinear Control*. 2017;27(6):894-914.
19. Xiong J, Lam J, Gao H, Ho DWC. On robust stabilization of Markovian jump systems with uncertain switching probabilities. *Automatica*. 2005;41(5):897-903.
20. Mudge TN, Al-Sadoun HB. A semi-Markov model for the performance of multiple-bus systems. *IEEE Trans Comput*. 1985;C-34(10):934-942.
21. Ji Y, Li Y, Wu W, Fu H, Qiao H. Mode-dependent event-triggered tracking control for uncertain semi-Markov systems with application to vertical take-off and landing helicopter. *Measur Control Lond Inst Measur Control*. 2020;53(5-6):954-961.
22. Wang J, Chen M, Shen H. Event-triggered dissipative filtering for networked semi-Markov jump systems and its applications in a mass-spring system model. *Nonlinear Dyn*. 2017;87:2741-2753.
23. Zhang L, Leng Y, Colaneri P. Stability and stabilization of discrete-time semi-Markov jump linear systems via semi-Markov kernel approach. *IEEE Trans Autom Control*. 2016;61(2):503-508.
24. Zhang L, Yang T, Colaneri P. Stability and stabilization of semi-Markov jump linear systems with exponentially modulated periodic distributions of sojourn time. *IEEE Trans Autom Control*. 2017;62(6):2870-2885.
25. Huang J, Shi Y. Stochastic stability and robust stabilization of semi-Markov jump linear systems. *Int J Robust Nonlinear Control*. 2013;23(18):2028-2843.
26. Ning Z, Zhang L, Mesbah A, Colaneri P. Stability analysis and stabilization of discrete-time non-homogeneous semi-Markov jump linear systems: a polytopic approach. *Automatica*. 2020;120:109080.
27. Jiang B, Karimi HR. Further criterion for stochastic stability analysis of semi-Markovian jump linear systems. *Int J Robust Nonlinear Control*. 2020;30(7):2689-2700.
28. Zhang Y, Lim CC, Liu F. Robust control synthesis for discrete-time uncertain semi-Markov jump systems. *Int J Syst Sci*. 2019;50(10):2042-2052.
29. Jafari S, Savla K. A principled approximation framework for optimal control of semi-Markov jump linear systems. *IEEE Trans Autom Control*. 2019;64(9):3616-3631.
30. Wang B, Zhu Q. Stability analysis of discrete-time semi-Markov jump linear systems. *IEEE Trans Autom Control*. 2020;65(12):5415-5421.
31. Ning Z, Zhang L, Colaneri P. Semi-Markov jump linear systems with incomplete sojourn and transition information: analysis and synthesis. *IEEE Trans Autom Control*. 2020;65(1):159-174.
32. Tian Y, Yan H, Zhang H, Zhan X, Peng Y. Dynamic output-feedback control of linear semi-Markov jump systems with incomplete semi-Markov kernel. *Automatica*. 2020;117:108997.

33. Sutton RS, Barto AG. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press; 2018.
34. Lewis FL, D V, Vamvoudakis K G. Reinforcement learning and feedback control: using natural decision methods to design optimal adaptive controllers. *Control Syst IEEE*. 2012;32(6):76-105.
35. Rizvi SA, Lin Z. Output feedback Q-learning control for the discrete-time linear quadratic regulator problem. *IEEE Trans Neural Netw Learn Syst*. 2019;30(5):1523-1536.
36. Al-Tamimi A, Lewis FL, Abu-Khalaf M. Model-free Q -learning designs for linear discrete-time zero-sum games with application to H_∞ control. *Automatica*. 2007;43(3):473-481.
37. Wang D, Qiao J, Cheng L. An approximate neuro-optimal solution of discounted guaranteed cost control design. *IEEE Trans Cybern*. 2020;1-10.
38. Wang D, Ha M, Qiao J. Data-driven iterative adaptive critic control towards an urban wastewater treatment plant. *IEEE Trans Ind Electron*. 2020;1.
39. Wang D, Ha M, Qiao J. Self-learning optimal regulation for discrete-time nonlinear systems under event-driven formulation. *IEEE Trans Autom Control*. 2020;65(3):1272-1279.
40. Jiang H, Zhang H, Luo Y, Wang J. Optimal tracking control for completely unknown nonlinear discrete-time Markov jump systems using data-based reinforcement learning method. *Neurocomputing*. 2016;194(19):176-182.
41. Chen Y, Wen J, Luan X, Liu F. Robust control for Markov jump linear systems with unknown transition probabilities Ċ an online temporal differences approach. *Trans Inst Meas Control*. 2020;42(15):3043-3051.
42. Samuelson PA. Interactions between the multiplier analysis and the principle of acceleration. *Rev Econ Stat*. 1939;21(2):75-78.
43. Blair S. Feedback control of a class of linear discrete systems with jump parameters and quadratic cost criteria. *Int J Control*. 1975;21(5):833-841.
44. Westerhoff FH. Samuelson's multiplier-accelerator model revisited. *Appl Econ Lett*. 2006;13(2):89-92.

How to cite this article: Chen Y, Wen J, Luan X, Liu F. H_∞ optimal control for semi-Markov jump linear systems via TP-free temporal difference (λ) learning. *Int J Robust Nonlinear Control*. 2021;1–12. https://doi.org/10.1002/rnc.5648