Parallel Optimal Tracking Control Schemes for Mode-Dependent Control of Coupled Markov Jump Systems via Integral RL Method

Kun Zhang, *Student Member, IEEE*, Hua-guang Zhang, *Fellow, IEEE*, Yuliang Cai, and Rong Su, *Senior Member, IEEE*

***Abstract*— This article is concerned with the optimal tracking control problem of the coupled Markov jump system (CMJS) by using the reinforcement learning (RL) technique. Based on the conventional optimal tracking architecture, an offline tracking iteration algorithm is first designed to solve the coupled algebraic Riccati equation that can hardly be solved by mathematical methods directly. To overcome the crucial requirements and existing shortcomings in the offline tracking method, a novel integral RL (IRL) tracking algorithm is first proposed for CMJS, which develops a transition-probability-free optimal tracking control scheme with a reconstructed augmented system and discounted cost function. Both the requirements of transition probability π_{ij} and system matrix A_i are avoided via the designed IRL algorithm. The stability and convergence of the novel schemes are proved by the Lyapunov theory, and the tracking objective is achieved as desired. Finally, we apply the designed algorithms in a fourth-order Markov jump control problem and the stochastic mass, spring, and damper system to track continuous sinusoidal waveforms, and the simulation results are provided to show the effectiveness and applicability.**

***Note to Practitioners*—In the practical engineering systems, many useful signals and interference vary randomly. Therefore, the tracking control of stochastic systems and dynamics, such as the Markovion, Itô's, Wiener, and Martingale processes, plays an important role in the modern industry. As a matter of fact, it is always desired to reduce the requirement of exact information and transition probability in the homogeneous Markovian process, which is very difficult to obtain accurate measurements. One way is integrating the adaptive reinforcement learning (RL) technique into the Markovian systems to learn this implicit information. However, a major restriction of the RL technique is that the control policy should be related to the finite performance index, which generally invalidates the optimal tracking solutions. In order to tackle this difficulty, by designing a novel parallel scheme via integral RL (IRL) technique, the solution of the coupled algebraic Riccati equation is solved, and the transition probability can be completely unknown during the learning process.**

***Index Terms*— Adaptive dynamic programming (ADP), Markov jump system, optimal control, stochastic stability, tracking control.**

I. Introduction

DURING the last decades, evolutionary computation and control optimization algorithms have received significantly increasing attention and have gradually become one key focus of intelligent control fields. Reinforcement learning (RL) methods can eliminate the unavoidable curse of dimensionality effectively in traditional dynamic programming theory [1]–[3]. Unlike some general control approaches [4], [5], the integral RL (IRL) technique can map a relationship between the optimal control policy and the system dynamics directly. Based on the essence and superiority of RL methods, the optimal control can minimize the value/cost function in performance index [6], [7]. A recent objective of the researchers in the field of control is to develop the IRL technique so as to find the optimal controllers available in industrial practice and engineering applications.

The optimal tracking control problem (OTCP) has become a key issue in aerocraft design and military applications, and many solutions have been put forward [8]–[10]. Adaptive dynamic programming (ADP) [11], [12] and some iteration RL methods [13]–[15], developed to overcome the difficulties caused by solving the nonlinear mathematics, are widely used to approximate the optimal policy. In the traditional ADP-based solution for tracking problems, the control law consists of the steady-state control input and the feedback control input [16]–[18]. Meanwhile, different from the traditional method, an augmented-system tracking control method was designed to seek out the optimal control under the discounted performance index, where the system dynamics can be either completely or partially unknown [19], [20]. However, only simple mathematical models in the above-mentioned literature were considered. Many factors, such as the Markovian processes, or other features often encountered in practical applications, were not discussed.

In most of the practical engineering systems, the complex Markovian dynamics play a key role [21]–[23], and the

Manuscript received October 5, 2019; accepted October 16, 2019. Date of publication November 12, 2019; date of current version July 2, 2020. This article was recommended for publication by Associate Editor B. Fidan and Editor Q. Zhao upon evaluation of the reviewers' comments. This work was supported in part by the National Natural Science Foundation of China under Grant 61627809, Grant 61433004, and Grant 61621004 and in part by the Liaoning Revitalization Talents Program under Grant XLYC1801005. *(Corresponding author: Huaguang Zhang.)*

K. Zhang and Y. Cai are with the School of Information Science and Engineering, Northeastern University, Shenyang 110819, China (e-mail: nukgnahz@163.com; caiyuliangfly@126.com).

H. Zhang is with the State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang 110819, China, and also with the School of Information Science and Engineering, Northeastern University, Shenyang 110819, China (e-mail: hgzhang@ieee.org).

R. Su is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: rsu@ntu.edu.sg).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASE.2019.2948431

stability and controllability of the Markov jump systems have received much attention for its powerful modeling capability on aerospace systems, power systems, physical systems, and manufacturing systems [24]–[26]. To analyze the stabilizability properties and the optimal control of the infinite time Markov jump systems, the necessary and sufficient conditions were established for stochastic linear systems in [27]. As one of the most important methods to solve the coupled algebraic Riccati equation (CARE), optimal preview control was developed in [28] and [29] and parallel algorithms were proposed to approximate the optimal solutions in [30]–[32]. For the Markov jump systems, the solution of CARE was equal to the coupled quadratic Lyapunov equations, and then, a convex programming and policy iteration methods were designed to decouple the complex CARE by using subsystems' transformation method [33]–[35]. Besides, the data-driven sliding mode and linear matrix inequalities methods were developed in [36] and [37] to deal with the Markov jump nonlinear systems, which found two tracking controls for the stochastic dynamics. The ADP-based schemes were also utilized for the Markov jump systems in [38] and [39], where the iterative method was proposed and employed to update the control policies. Unfortunately, there are few learning algorithms designed for the tracking control of stochastic systems, and the multimodal switching in the Markovian process is still a thorny problem that can hardly be solved by the mathematical analytic methods directly.

In this article, two optimal tracking control algorithms are proposed for the coupled Markov jump dynamics by using offline and online parallel learning techniques, respectively. The main contributions are summarized as follows.

1) To the best of our knowledge, there are no existing mathematical approaches to solve the Markovian jump OTCP directly, and the designed control algorithms in this article are the first of their kind for the optimal tracking control solution of the Markovian jump systems.
2) Compared with some existing methods to solve CARE problem [14], [31], [32], the system information can be partially unknown in the novel IRL algorithm, where the system matrices A_i, F and stationary transition probability π_{ij} are avoided during the solving process.
3) The stochastic stability and convergence of the developed algorithms for the Markovian jump OTCP are first proved by theorems that guarantee the correctness of algorithms. Besides, the validity of the designed methods is confirmed by simulations effectively.

The rest of this article is organized as follows. In Section II, the problem formulation is given. The offline optimal tracking control algorithm is presented in Section III, where the convergence and stochastic stability are obtained. Based on the new constructed augmented system and the corresponding discounted cost function, a novel transition-probability-free optimal tracking control algorithm via IRL technique is first proposed in Section IV, and both the convergence and stochastic stability of the novel algorithm are proved and guaranteed by theorems in this section. These tracking control algorithms are utilized in simulations, and the effectiveness is demonstrated in Section V. Finally, a brief conclusion is drawn in Section VI.

II. Problem Formulation

In this article, let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space, where Ω is the sample space, $\mathcal{F}$ is the σ-algebra of events, and $\mathcal{P}$ is the probability measure defined on $\mathcal{F}$. We consider a class of continuous-time Markovian jump systems described as

$$\dot{x}(t) = A(r(t))x(t) + B(r(t))u(t), r(t) \in \Im \tag{1}$$

where $x(t) \in \mathbb{R}^n$ is the system state, both $A(r(t))$ and $B(r(t))$ are known mode-dependent matrices, $\{r(t)\}$ is a homogeneous Markovian process with right continuous trajectories on finite discrete state space $\Im = \{1, 2, \ldots, N\}$, and $u(t) \in \mathbb{R}^m$ is the control input.

The stationary transition probability of the transition-probability-rate matrix $\Pi \triangleq [\pi_{ij}]_{N\times N}$ in the homogeneous Markovian process is

$$\Pr\{r(t+\Delta) = j | r(t) = i\} := \begin{cases} \pi_{ij}\Delta + o(\Delta), & i \neq j \\ 1 - \pi_i \Delta + o(\Delta), & i = j \end{cases} \tag{2}$$

where $\Delta > 0$, $\lim_{\Delta\to 0} o(\Delta)\Delta^{-1} = 0$, $\pi_{ij} \geq 0$ is the probability rate between modes i at time t and j at time $t + \Delta$, for $i \neq j$; $i, j \in \Im$ and $\forall i \in \Im$, and $\pi_i := -\pi_{ii} = \sum_{j=1, j\neq i}^{N} \pi_{ij}$.

To stabilize the stochastic system (1) in the Markovian process (2), the objective is to find a set of mode-dependent control policy $\varphi(t, x(t), r(t))$ as

$$u(t, r(t)) \triangleq \varphi(t, x(t), r(t)) = K(r(t))x(t)$$
$$\varphi : [0, \infty) \times \mathbb{R}^n \times \Im \to \mathbb{R}^m. \tag{3}$$

Definition 1 ([39]): The continuous-time Markovian jump system (1) is stochastically stable if, for any finite $x(0) = x_0 \in \mathbb{R}^n$ and $r(0) = r_0 \in \Im$, there exists the linear feedback control $u(t, r(t))$ satisfying

$$\lim_{T\to\infty} E\left\{\int_0^T x^{\mathrm{T}}(t, u)x(t, u)dt | x_0, r_0\right\} \leq x_0^{\mathrm{T}} M x_0 \tag{4}$$

where M is a symmetric positive matrix.

For the optimal control, let the infinite horizon performance index with respect to system (1) be defined as

$$V(t, x(t), r(t)) = E\left\{\int_0^\infty [x^{\mathrm{T}}(t)Q(r(t))x(t) + u^{\mathrm{T}}(t, r(t))R(r(t))u(t, r(t))]dt | x_0, r_0\right\} \tag{5}$$

where $Q(r(t)) > 0$ and $R(r(t)) > 0$.

Assumption 1 ([27]): We assume that the system (1) is stochastically controllable, the values of $r(t)$ and $x(t)$ are available at time t exactly, and the expectation in (5) is over the joint process $\{x(t), r(t)\}$. The modes jump with a minimum residence time, such as Δ, in the homogeneous Markovian process.

For convenience, we will denote $A(r(t))$, $B(r(t))$, $Q(r(t))$, $R(r(t))$, $u(t, r(t))$, and $K(r(t))$ for $r(t) = i$, $i \in \Im$, by A_i, B_i, Q_i, R_i, $u_i(t)$, and K_i in the subsequent development.

Assumption 2: The continuous-time Markovian jump stochastic system (1) can be stochastically stable under the set optimal control policies $u_i^*(t)$ for all $i \in \Im$, and the pairs $(A_i, B_i, \sqrt{Q_i})$, $i \in \Im$ are stochastically detectable.

Definition 2: A mode-dependent control policy $u_i(t) = \varphi(t, x(t), i)$ with $\varphi(0, x(0), r(0))$ called to be a mode-dependent admissible control for such system (1), if the control policy $\varphi(t, x(t), i) \in \mathbb{R}^m$ is continuous, can stochastically stabilize the system (1) and make the related mode-dependent cost function $V(t, x(t), r(t) = i)$ finite.

Then, the infinite horizon optimal control problem of the Markovian jump system (1) is transformed to find the following set of mode-dependent admissible control policies:

$$u_i^*(t) \triangleq \underset{u_i \in \Psi(\Omega)}{\operatorname{argmin}} V_i(x(t), u_i(t)) \quad \forall i \in \Im \tag{6}$$

where $V_i(x(t), u_i(t)) = V(t, x(t), r(t) = i)$. The optimal cost function is, thus, obtained as $V_i^*(x(t), u_i^*(t))$.

Thus, for a Markovian jump system, the optimal control solution is equivalent to solve the CARE [13], [31], [32]. Suppose the stochastic CARE has a unique optimal solution sequence $P = (P_1, \ldots, P_i, \ldots, P_N)$, $i \in \Im$, $P_i > 0$, satisfying

$$[A_i + B_i K_i]^{\mathrm{T}} P_i + P_i [A_i + B_i K_i] = -K_i^{\mathrm{T}} R_i K_i - \sum_{j=1}^{N} \pi_{ij} P_j - Q_i \tag{7}$$

then the optimal control policy yields

$$u_i(t) = K_i x(t) = -R_i^{-1} B_i^{\mathrm{T}} P_i x(t) \tag{8}$$

where K_i is the optimal control gain for $i \in \Im$.

To solve the CARE of optimal tracking control solution with the continuous-time Markovian jump system (1), an offline parallel control scheme and a novel IRL tracking control algorithm are designed in this article.

III. Transition-Probability-Based Offline Parallel Tracking Control Algorithm for Markov Jump Systems

In this section, a transition-probability-based offline method to solve the CARE of OTCP is presented. It is assumed that the reference trajectory is generated from a command generator, and the cost function is quadratic in the tracking error.

Let the desired reference trajectory take the following form:

$$\dot{x}_d(t) = F x_d(t) \tag{9}$$

where $x_d \in \mathbb{R}^n$ is the desired state and $F \in \mathbb{R}^{n \times n}$ is a constant matrix.

Remark 1: Note that the dynamics from the command generator are not necessarily asymptotic stable and can use in many different applications, such as the sinusoidal waveforms, damped sinusoids, and the ramp in satellite antenna pointing. This is different from the traditional tracking ADP methods, where the desired trajectory needs to be stable and converging to zero.

The objective of the OTCP is to find a control policy to make the system states track the reference trajectory while minimizing the cost function. Then, the Markov jump tracking error dynamic becomes

$$\dot{e}(t) = A_i x(t) - F x_d(t) + B_i u_i(t) \tag{10}$$

where $e(t) = x(t) - x_d(t)$ is the tracking error.

To follow the conventional optimal tracking architecture, we can rewrite the reference trajectory as follows:

$$\dot{x}_d(t) = A_i x_d(t) + B_i u_i^c(t) \tag{11}$$

where $u_i^c(t)$ is the conventional steady-state control input taking the following form:

$$u_i^c(t) = B_i^{+} [F x_d(t) - A_i x_d(t)] \tag{12}$$

where B_i^{+} is the generalized inverse of B_i.

Define a new control input as $u_i^d(t) = u_i(t) - u_i^c(t)$, which is the feedback control with respect to the tracking error $e(t)$. Hence, the corresponding tracking cost function becomes

$$V(t, e(t), r(t)) = E\left\{ \int_t^{\infty} \left[e^{\mathrm{T}}(\tau) Q_i e(\tau) + u_i^{d\mathrm{T}}(\tau) R_i u_i^d(\tau) \right] d\tau \,|\, e(t), r(t) \right\} \tag{13}$$

where $u_i^d(t) = -R_i^{-1} B_i^{\mathrm{T}} P_i e(t)$.

By using the designed control method, the unstable-reference tracking control problem is transformed into an optimal regulation problem, where the optimal feedback control minimizes the cost function (13). Based on the conventional optimal tracking architecture, a transition-probability-based offline tracking scheme for the Markov Jump systems is proposed in Algorithm 1.

Theorem 1: The Markovian jump tracking dynamic (10) is stochastically stable by using Algorithm 1, and the iteration solution in the tracking algorithm converges to the solution of the CARE if there exists a feedback control gain $\mathcal{K}_i$ such that the matrix $Q_i + \mathcal{K}_i^{\mathrm{T}} R_i \mathcal{K}_i$ is positive definite for each $i \in \Im$.

Proof: [Stability] According to the designed offline tracking architecture, we simplify the tracking error dynamics (10) by using the steady-state control (16a) as

$$\begin{aligned} \dot{e}(t) &= A_i x(t) - F x_d(t) + B_i (u_i^c(t) + u_i^d(t)) \\ &= A_i x(t) - F x_d(t) + [F - A_i] x_d(t) + B_i u_i^d(t) \\ &= A_i e(t) + B_i u_i^d(t). \end{aligned} \tag{17}$$

Inserting the feedback control law (16b) into the tracking error dynamic (17), the closed-loop system becomes $\dot{e}(t) = (A_i + B_i \mathcal{K}_i) e(t)$. Then, select the stochastic quadratic Lyapunov function as $V(e(t), r(t) = i) = V(e, i) = e^{\mathrm{T}} \mathcal{P}_i e$.

Considering the weak infinitesimal operator $\mathfrak{L}$ of the joint process $\{(r(t), e(t)), t \in [0, T]\}$ as the natural stochastic analog of the deterministic derivative and the function space

Algorithm 1 Transition-Probability-Based Off-Line Optimal Tracking Control Algorithm for the Markov Jump Systems

Initialization Select an initial matrix sequence $P(0) = (P_1(0), \cdots, P_N(0))$, initial admissible control gain $K_i(0)$ and $u_i^c(0)$, $i \in \Im$. Consider the tracking dynamic (10) and let the iteration be conducted at any time t and $k = 0$.

Procedure

1: **while** $\max_{i\in\Im} \|P_i(k+1) - P_i(k)\| \geq \epsilon$, $\epsilon > 0$ **do**
2: **for** i=i:N **do**
3: Solve the following continuous-time Lyapunov equation for $P_i(k+1)$ as

$$\begin{aligned}&[\mathcal{A}_i + B_i K_i(k)]^{\mathrm{T}} P_i(k+1) \\ &\quad + P_i(k+1)[\mathcal{A}_i + B_i K_i(k)] \\ &= -K_i^{\mathrm{T}}(k) R_i K_i(k) - \sum_{j=1 j\neq i}^{N} \pi_{ij} P_j(k) - Q_i\end{aligned} \tag{14}$$

where $\mathcal{A}_i = A_i + (\pi_{ii}/2) I_n$ and I_n is a n-dimensional identity matrix.

4: Update the feedback control gain by

$$K_i(k+1) = -R_i^{-1} B_i^{\mathrm{T}} P_i(k+1). \tag{15}$$

5: Set $\mathcal{P}_i = P_i(k+1)$ and $\mathcal{K}_i = K_i(k+1)$
6: **end for**
7: **end while** ▷ the iteration is finished
8: Use the steady-state control input and the feedback control input by

$$u_i^c(t) = B_i^{+}[F - A_i] x_d(t) \tag{16a}$$

$$u_i^d(t) = \mathcal{K}_i e(t) = \mathcal{K}_i (x(t) - x_d(t)). \tag{16b}$$

End Procedure

$[0, T] \times \Im \times \mathbb{R}^n$ as the domain of $\mathfrak{L}$, one has

$$\begin{aligned}\mathfrak{L}V(e,i) &\triangleq \lim_{\triangle\to 0} \frac{1}{\triangle}[E\{V(e(t+\triangle), r(t+\triangle))|e(t), r(t) = i\} \\ &\qquad - V(e(t), r(t) = i)] \\ &= e^{\mathrm{T}}(t) \Bigg[(A_i + B_i\mathcal{K}_i)^{\mathrm{T}}\mathcal{P}_i + \mathcal{P}_i(A_i + B_i\mathcal{K}_i) \\ &\qquad + \sum_{j=1}^{N} \pi_{ij}\mathcal{P}_j \Bigg] e(t) \\ &= e^{\mathrm{T}}(t) \Bigg[(\mathcal{A}_i + B_i\mathcal{K}_i)^{\mathrm{T}}\mathcal{P}_i + \mathcal{P}_i(\mathcal{A}_i + B_i\mathcal{K}_i) \\ &\qquad + \sum_{j=1 j\neq i}^{N} \pi_{ij}\mathcal{P}_j \Bigg] e(t) \\ &= -e^{\mathrm{T}}(t)\big[Q_i + \mathcal{K}_i^{\mathrm{T}} R_i \mathcal{K}_i\big] e(t) \leq 0.\end{aligned} \tag{18}$$

Thus, it can be concluded that $\mathfrak{L}V(e,i) \leq -\beta V(e,i)$, where $\beta = \min_{i\in\Im}(\lambda_{\min}(Q_i + \mathcal{K}_i^{\mathrm{T}} R_i \mathcal{K}_i)/\lambda_{\max}(\mathcal{P}_i)) > 0$, $\lambda_{\min}(\cdot)$ is the minimum eigenvalue, and $\lambda_{\max}(\cdot)$ is the maximum eigenvalue.

By using Dynkin's formula and the Gronwell–Bellman lemma for all $i \in \Im$ and $e_0 = e(0)$, it yields

$$E\{V(e(t), i)|e_0, r_0 = i\} \leq \exp(-\beta t) e_0^{\mathrm{T}} \mathcal{P}_i e_0 \tag{19}$$

which is integrated as the following equation:

$$\begin{aligned}&E\left\{\int_0^T e^{\mathrm{T}}(t)\mathcal{P}_i e(t) dt | e_0, r_0 = i\right\} \\ &\qquad \leq \left(\int_0^T \exp(-\beta t) dt\right) e_0^{\mathrm{T}} \mathcal{P}_i e_0\end{aligned} \tag{20}$$

where $\exp(\cdot)$ is the exponential function.

Tacking limit as $T \to \infty$, one has

$$\lim_{\mathrm{T}\to\infty} E\left\{\int_0^T e^{\mathrm{T}}(t) e(t) dt | e_0, r_0 = i\right\} \leq e_0^{\mathrm{T}} M e_0 \tag{21}$$

where $M = \max_{i\in\Im}(\mathcal{P}_i/(\beta||\mathcal{P}_i||))$. This implies that (4) holds.

Convergence: Under the aforementioned assumptions, by using the designed steady-state control (16a) in solution procedure, the Markovian jump tracking solution reduces to the solution of stochastic CARE (7) and the optimal feedback control policy (8). Define the mapping on a Banach space as $\mathcal{F}(P_i) \triangleq [A_i + B_i K_i]^{\mathrm{T}} P_i + P_i[A_i + B_i K_i] + K_i^{\mathrm{T}} R_i K_i + \sum_{j=1}^{N} \pi_{ij} P_j + Q_i = 0$ and use Fréchet derivative; hence, steps 2–6 in the iteration procedure will be equate to Newton's method by using Kantorovtich's theorem [40]. Then, the matrix $\mathcal{P}_i$ and control gain $\mathcal{K}_i$, $i \in \Im$, converge to the optimal solutions, and the so-called quadratic Lyapunov equations are solved. The proof is, thus, completed. ■

Remark 2: Note that there is almost no way to solve the CARE of the coupled Markov jump system (CMJS) by mathematical methods directly. By adopting the exact information of transition probability and system matrices, the designed offline tracking control algorithm decouples the CARE problem and obtains the optimal solution successfully. Both the stochastic stability and convergence of the Markov jump tracking dynamics via the designed Algorithm 1 are guaranteed, and the tracking errors will decrease to zero by using the achieved control.

Remark 3: In the offline tracking control algorithm, some requirements or shortcomings are crucial: 1) the control policy consists of steady-state control input and feedback control, where the steady-state control input is assumed that it can be derived as (12) directly and 2) the exact information of transition probability in the homogeneous Markovian process and the system matrices are essential and necessary for every iteration computation until the termination is obtained.

To overcome these crucial requirements and shortcomings in iteration computation above, a novel transition-probability-free optimal tracking control algorithm via the IRL technique is proposed in Section IV to solve the optimal tracking control in real time, which better meets the industry requirements from practical applications.

IV. Transition-Probability-Free IRL Optimal Tracking Control Algorithm for Markov Jump Systems

According to the offline algorithm, the exact information of the system matrices and the transitional probability are

required to solve the complicated CARE. However, in practical applications, this information can hardly be observed in the procedures, and there is almost no way to measure the transition probability precisely. To obtain the OTCP solution of the Markovian jump systems without the information of transition probability in the homogeneous Markovian process, a novel IRL-based tracking control algorithm is designed in this section.

Based on the continuous Markov jump system (1) and the desired trajectory dynamics (9), respectively, there is

$$\begin{aligned}\dot{e}(t) &= \dot{x}(t) - \dot{x}_d(t)\\ &= A_i x(t) - A_i x_d(t) + A_i x_d(t) - F x_d(t) + B_i u_i(t)\\ &= A_i e(t) + (A_i - F)x_d(t) + B_i u_i(t). \end{aligned} \tag{22}$$

Hence, we construct an augmented system by

$$\dot{X}(t) = \begin{bmatrix} A_i & A_i - F \\ \mathbf{0} & F \end{bmatrix} X(t) + \begin{bmatrix} B_i \\ \mathbf{0} \end{bmatrix} u_i(t) \equiv \bar{A}_i X + \bar{B}_i u_i \tag{23}$$

where the state is $X(t) = \left[e^{\mathrm{T}}(t), x_d^{\mathrm{T}}(t)\right]^{\mathrm{T}} \in \mathbb{R}^{2n}$, $\mathbf{0}$ is a zero matrix with an appropriate dimension, and the fixed feedback control is $u_i(t) = \bar{K}_i X(t) \in \mathbb{R}^m$.

Let the corresponding cost function be defined as

$$\begin{aligned} V(X(t), r(t)) = E\Bigg\{ \int_t^\infty \eta^{\lambda(\tau - t)} \big[& X^{\mathrm{T}}(\tau) \bar{Q}_i X(\tau) \\ & + u_i^{\mathrm{T}}(\tau) R_i u_i(\tau) \big] d\tau \,|\, X(t), r(t) \Bigg\} \end{aligned} \tag{24}$$

where $0 < \eta < 1$ is the discount factor, $\lambda > 0$ is a chosen parameter, and $\bar{Q}_i$ is a positive definite matrix. Hence, the coupled algebraic Riccati equation yields

$$\begin{aligned} &[\bar{A}_i + \bar{B}_i \bar{K}_i]^{\mathrm{T}} \bar{P}_i + \bar{P}_i [\bar{A}_i + \bar{B}_i \bar{K}_i] \\ &\quad = -\bar{K}_i^{\mathrm{T}} R_i \bar{K}_i - \sum_{j=1}^{N} \pi_{ij} \bar{P}_j - \lambda \ln \eta \bar{P}_i - \bar{Q}_i. \end{aligned} \tag{25}$$

Lemma 1: Considering the constructed augmented system (23), then the stochastic cost function (24) with control $u_i(t)$ can be written as the quadratic form $V(X(t), r(t) = i) = \left[e^{\mathrm{T}}(t), x_d^{\mathrm{T}}(t)\right] \bar{\mathcal{P}}_i \left[e^{\mathrm{T}}(t), x_d^{\mathrm{T}}(t)\right]^{\mathrm{T}} > 0$ for some symmetric $\bar{\mathcal{P}}_i$.

Proof: According to the fixed admissible feedback control $u_i(t)$ and the expectation of solution of linear differential equation (23), the cost function becomes

$$\begin{aligned} &V(X(t), r(t) = i) \\ &= E\Bigg\{ \int_t^\infty \eta^{\lambda(\tau - t)} [X^{\mathrm{T}}(\tau) \bar{Q}_i X(\tau) + [\bar{K}_i X(\tau)]^{\mathrm{T}} \\ &\qquad \downarrow \tau_1 = \tau - t \qquad\qquad \times R_i [\bar{K}_i X(\tau)]] d\tau \,|\, X(t), r(t) \Bigg\} \\ &= E\Bigg\{ \int_0^\infty \eta^{\lambda\tau} \big[X^{\mathrm{T}}(\tau + t) \big[\bar{Q}_i + \bar{K}_i^{\mathrm{T}} R_i \bar{K}_i \big] \\ &\qquad\qquad \times X(\tau + t) \big] \big] d\tau \,|\, X(t), r(t) \Bigg\} \\ &= E\Bigg\{ \int_0^\infty \eta^{\lambda\tau} \big[X^{\mathrm{T}}(t) \exp^{\mathrm{T}}(\bar{L}_i \tau) \big[\bar{Q}_i + \bar{K}_i^{\mathrm{T}} R_i \bar{K}_i \big] \\ &\qquad\qquad \times \exp(\bar{L}_i \tau) X(t) \big] d\tau \,|\, X(t), r(t) \Bigg\}. \end{aligned} \tag{26}$$

$\dot{X}(t) = \bar{L}_i X(t) \Rightarrow X(t-t) = e^{\bar{L}_i t} X(t)$
$X(\tau + t) = e^{\bar{L}_i(\tau + t)} X(0) = e^{\bar{L}_i \tau} X(t)$

Algorithm 2 Transition-Probability-Free IRL Optimal Tracking Control Algorithm for the Markov Jump Systems

Initialization

Select an initial sequence $\bar{P}(0) = (\bar{P}_1(0), \cdots, \bar{P}_N(0))$, initial admissible feedback control gain $\mathcal{K}_i = \bar{K}_i(0)$, $\forall i \in \Im$. Let $\triangle$ be small enough and the tracking procedure be propagating along with time $t \in [0, +\infty)$.

Procedure

1: **while** $t <$ T, T is a selected terminal time **do**
2: **for** i=1:N **do**
3: Solve the following decoupled continuous-time Lyapunov equation as

$$\begin{aligned} &X^{\mathrm{T}}(t) \bar{P}_i(k) X(t) - X^{\mathrm{T}}(t + \triangle) \bar{P}_i(k) X(t + \triangle) \\ &\quad = \int_t^{t+\triangle} X^{\mathrm{T}}(\varsigma) \big[\bar{Q}_i + \lambda \ln \eta \bar{P}_i(k-1) \\ &\qquad + \bar{K}_i(k)^T R_i \bar{K}_i(k) \big] X(\varsigma) d\varsigma \end{aligned} \tag{28}$$

4: Update the feedback control gain and the decoupled parameter by

$$\bar{K}_i(k+1) = -R_i^{-1} \bar{B}_i^{\mathrm{T}} \bar{P}_i(k) \tag{29}$$

5: Set the positive matrix $\bar{\mathcal{P}}_i = \bar{P}_i(k)$ and feedback control gain $\mathcal{K}_i = \bar{K}_i(k+1)$
6: **end for**
7: Use the feedback control as

$$u_i(k+1, t) = \mathcal{K}_i X(t), i \in \Im \tag{30}$$

8: **end while**

End Procedure

Extracting the state from integration, it has

$$\begin{aligned} &V(X(t), r(t) = i) \\ &\quad = X^{\mathrm{T}}(t) \Bigg(E\Bigg\{ \int_0^\infty \eta^{\lambda\tau} \big[\exp^{\mathrm{T}}(\bar{L}_i \tau) \big[\bar{Q}_i + \bar{K}_i^{\mathrm{T}} R_i \bar{K}_i \big] \\ &\qquad\qquad \times \exp(\bar{L}_i \tau) \big] d\tau \,|\, X(t), r(t) \Bigg\} \Bigg) X(t) \\ &\quad = X^{\mathrm{T}}(t) \bar{\mathcal{P}}_i X(t) \end{aligned} \tag{27}$$

where $\bar{L}_i = \bar{A}_i + \bar{B}_i \bar{K}_i$, $\bar{\mathcal{P}}_i = E\{\int_0^\infty \eta^{\lambda\tau} [\exp^{\mathrm{T}}(\bar{L}_i \tau) [\bar{Q}_i + \bar{K}_i^{\mathrm{T}} R_i \bar{K}_i] \exp(\bar{L}_i \tau)] d\tau \,|\, X(t), r(t)\}$, and $\exp(\cdot) : \mathbb{R}^p \to \mathbb{R}^p$ is a continuous one-to-one function as $\exp([v_1, \ldots, v_p]) = [\exp(v_1), \ldots, \exp(v_p)]$. This completes the proof. ■

According to the augmented tracking system (23) and cost function (24), the transition-probability-free IRL parallel tracking control algorithm above is developed to solve the OTCP.

Remark 4: Compared with the offline control methods for the Markov jump systems, the requirements of stationary transition probability π_{ij} and system matrices A_i are avoided by using the iteration equations (28) and (29) in the solving procedure, and the novel transition-probability-free design can be applied in the partially unknown CMJS. Instead of the measured information in the homogeneous Markovian process, the system matrices and transition probability are successfully eliminated by the novel algorithm.

As the optimal tracking control is achieved via the novel IRL method, the system dynamics can be guaranteed stochastically stable and the tracking errors get converged to zero as desired. The random parameters here will affect the learning or computing time in the learning process of the new algorithm. Both the stability and convergence of the novel transition-probability-free IRL algorithm are obtained by the following theorems.

Theorem 2 (Stability): Under the aforementioned assumptions and the initial stabilizing sequence $\bar{P}(0)$, the tracking error $e(t)$ in the augmented system (23) can be guaranteed to be stochastically stable during the iteration procedure and converge to zero as desired by using the proposed IRL tracking control algorithm (see Algorithm 2).

Proof: To analyze Algorithm 2 for the parallel tracking control problem, the solution procedure is presented as follows.

Stability With the k-Step Feedback Control: Considering the augmented closed-loop system dynamics $\dot{X}(t) = (\bar{A}_i + \bar{B}_i\bar{K}_i(k))X(t)$ in any kth iteration (28), there exists a unique set solution of the quadratic Lyapunov equation $\bar{P}(k) = (\bar{P}_1(k), \ldots, \bar{P}_i(k), \ldots, \bar{P}_N(k))$, $i \in \Im$, $\bar{P}_i(k) > 0$ satisfying (28). Selecting parameters λ and η satisfying $\bar{Q}_i + \lambda \ln \eta \bar{P}_i(k) \geq \underline{Q}_i > 0\,\forall k$, and since the quadratic stochastic Lyapunov function is $V_k(X(t), r(t) = i) = X^{\mathrm{T}}(t)\bar{P}_i(k)X(t)$ in the kth iteration, the feedback control law is $u_i(k,t) = \bar{K}_i(k)X(t)$, and the weak infinitesimal operator $\mathfrak{L}$ defined as in Theorem 1, it becomes

$$\begin{aligned}
&\mathfrak{L}V_k(X,i)\\
&\triangleq \lim_{\triangle\to 0}\frac{1}{\triangle}[E\{V_k(X(t+\triangle), r(t+\triangle))|X(t), r(t)=i\}\\
&\qquad - V_k(X(t), r(t)=i)]\\
&= X^{\mathrm{T}}(t)\Bigg[(\bar{\mathcal{A}}_i + \bar{B}_i\bar{K}_i(k))^{\mathrm{T}}\bar{P}_i(k)\\
&\qquad + \bar{P}_i(k)(\bar{\mathcal{A}}_i + \bar{B}_i\bar{K}_i(k)) + \sum_{j=1 j\neq i}^{N}\pi_{ij}\bar{P}_j(k)\Bigg]\\
&\quad \times X(t)\\
&= -X^{\mathrm{T}}(t)[\bar{Q}_i + \lambda\ln\eta\bar{P}_i(k-1) + \bar{K}_i^{T}(k)R_i\bar{K}_i(k)]X(t)\\
&\leq -X^{\mathrm{T}}(t)\big[\mathcal{Q}_i + \bar{K}_i^{T}(k)R_i\bar{K}_i(k)\big]X(t) \leq 0 \qquad (31)
\end{aligned}$$

where $\bar{\mathcal{A}}_i = \bar{A}_i + ((\pi_{ii}/2)I_{2n})$ and I_{2n} is a $2n$-dimensional identity matrix.

Then, multiplying (31) by $\triangle$ has

$$\begin{aligned}
&\lim_{\triangle\to 0}[E\{V_k(X(t+\triangle), r(t+\triangle))|X(t), r(t)=i\}\\
&\quad - V_k(X(t), r(t)=i)]\\
&= \lim_{\triangle\to 0}\Bigg(\sum_{j=1 j\neq i}^{N}\triangle\pi_{ij}[X^{\mathrm{T}}(t+\triangle)\bar{P}_j(k)X(t+\triangle)\\
&\qquad - X^{\mathrm{T}}(t)\bar{P}_i(k)X(t)] + \left(1 - \sum_{j=1 j\neq i}^{N}\triangle\pi_{ij}\right)\\
&\qquad \times [X^{\mathrm{T}}(t+\triangle)\bar{P}_i(k)X(t+\triangle) - X^{\mathrm{T}}(t)\bar{P}_i(k)X(t)]\Bigg)\\
&\triangleq [X^{\mathrm{T}}(t+\triangle)\bar{P}_i(k)X(t+\triangle) - X^{\mathrm{T}}(t)\bar{P}_i(k)X(t)]\\
&= \int_t^{t+\triangle} X^{\mathrm{T}}(\varsigma)\Bigg[(\bar{\mathcal{A}}_i + \bar{B}_i\bar{K}_i(k))^{\mathrm{T}}\bar{P}_i(k)\\
&\qquad + \bar{P}_i(k)(\bar{\mathcal{A}}_i + \bar{B}_i\bar{K}_i(k))\\
&\qquad + \sum_{j=1 j\neq i}^{N}\pi_{ij}\bar{P}_j(k)\Bigg]X(\varsigma)d\varsigma\\
&\leq -\int_t^{t+\triangle} X^{\mathrm{T}}(\varsigma)\big[\mathcal{Q}_i + \bar{K}_i^{\mathrm{T}}(k)\bar{R}_i\bar{K}_i(k)\big]X(\varsigma)d\varsigma \qquad (32)
\end{aligned}$$

which indicates (28), and the transition probability information in the complicated CARE gets avoided.

Thus, the stochastic control solution of complicated CARE can be transformed to be N decoupled Lyapunov equations as $\triangle \to 0$. By using the Dynkins formula and the Gronwell–Bellman lemma as in Theorem 1, (31) provides that the closed-loop system $\dot{X}(t) = [\bar{A}_i + \bar{B}_i\bar{K}_i(k)]X(t)$ is stochastically stable.

Stability With the (k + 1)-Step Feedback Control: As the feedback control law $u_i(k,t)$, $i \in \Im$, mentioned earlier is stabilizing the dynamics with $V_k(X,i)$, the updating law (29) is used for tuning the control gain to $\bar{K}_i(k+1)$. Considering the quadratic stochasitic Lyapunov function $V_k(X(t), r(t) = i) = X^{\mathrm{T}}(t)\bar{P}_i(k)X(t)$, the weak infinitesimal operator $\mathfrak{L}$ along the trajectory generated by feedback control law $u_i(k+1,t)$, and using the updating law (29), one obtains

$$\begin{aligned}
&\mathfrak{L}V_k(X,i)\\
&\triangleq X^{\mathrm{T}}(t)\Bigg[(\bar{\mathcal{A}}_i + \bar{B}_i\bar{K}_i(k+1))^{\mathrm{T}}\bar{P}_i(k)\\
&\qquad + \bar{P}_i(k)(\bar{\mathcal{A}}_i + \bar{B}_i\bar{K}_i(k+1)) + \sum_{j=1 j\neq i}^{N}\pi_{ij}\bar{P}_j(k)\Bigg]\\
&\quad \times X(t)\\
&= X^{\mathrm{T}}(t)\Bigg[(\bar{\mathcal{A}}_i + \bar{B}_i\bar{K}_i(k))^{T}\bar{P}_i(k)\\
&\qquad + \bar{P}_i(k)(\bar{\mathcal{A}}_i + \bar{B}_i\bar{K}_i(k)) + \sum_{j=1 j\neq i}^{N}\pi_{ij}\bar{P}_j(k)\Bigg]X(t)\\
&\quad + X^{\mathrm{T}}(t)\big[(\bar{K}_i(k+1) - \bar{K}_i(k))^{\mathrm{T}}\bar{B}_i^{\mathrm{T}}\bar{P}_i(k)\\
&\qquad + \bar{P}_i(k)\bar{B}_i(\bar{K}_i(k+1) - \bar{K}_i(k))\big]X(t)\\
&= -X^{\mathrm{T}}(t)\big[\bar{Q}_i + \lambda\ln\eta\bar{P}_i(k-1) + \bar{K}_i^{T}(k)R_i\bar{K}_i(k)\big]X(t)\\
&\quad + X^{\mathrm{T}}(t)\big[(\bar{K}_i(k) - \bar{K}_i(k+1))^{\mathrm{T}}R_i\bar{K}_i(k+1)\\
&\qquad + \bar{K}_i^{\mathrm{T}}(k+1)R_i(\bar{K}_i(k) - \bar{K}_i(k+1))\big]X(t)\\
&\leq -X^{\mathrm{T}}(t)[\mathcal{Q}_i + \bar{K}_i^{T}(k)R_i\bar{K}_i(k)]X(t) - X^{\mathrm{T}}(t)\\
&\quad \times\big[(\bar{K}_i(k+1) - \bar{K}_i(k))^{\mathrm{T}}R_i(\bar{K}_i(k+1) - \bar{K}_i(k))\\
&\qquad + \bar{K}_i^{\mathrm{T}}(k+1)R_i\bar{K}_i(k+1) - \bar{K}_i^{\mathrm{T}}(k)R_i\bar{K}_i(k)\big]\\
&\quad \times X(t). \qquad (33)
\end{aligned}$$

Using (31) for the first quadratic term, one obtains

$$\begin{aligned}&\mathfrak{L}V_k(X,i)\\&\le -X^{\mathrm{T}}(t)\big[\mathcal{Q}_i+\bar{K}_i^{\mathrm{T}}(k+1)R_i\bar{K}_i(k+1)\\&\qquad+[\bar{K}_i(k+1)-\bar{K}_i(k)]^{\mathrm{T}}R_i[\bar{K}_i(k+1)-\bar{K}_i(k)]\big]\\&\quad\times X(t)\le 0.\end{aligned}\tag{34}$$

Hence, the updated control policy $u_i(k+1,t)$ with the control gain $\bar{K}_i(k+1)$, $i\in\Im$, stochastically stabilizes the tracking dynamic, and the tracking objective has been guaranteed achieved in the online parallel tracking process. The proof is, thus, completed. ■

Note that the information of stationary transition probability π_{ij} in the homogeneous Markovian process and the system matrix A_i are avoided by steps 2–6 of the proposed IRL tracking control algorithm.

Theorem 3 (Convergence): Let us consider the sequences $V_k(X,i)$ and $u_i(k,t)$, $i\in\Im$, obtained by (28) and (29), respectively. If $V_0(X,i)=0$ and the optimal cost function $V^*(X,i)>0\in C^1$ is smooth on a compact domain of validity Ω^*, then it follows that $V_k(X,i)\in C^1$ is a nondecreasing sequence and $V_{k+1}(X,i)\ge V_k(X,i)\ge 0\ \forall X\in\Omega^*$. Moreover, the sequence $V_k(X,i)$ converges to the solution of the CARE as $k\to\infty$, $V_k(X,i)\to V^*(X,i)$, and $u_i(k,t)\to u_i^*(t)$, where $u_i^*(t)$ is the optimal feedback control law.

Proof: According to Theorem 2, the stability of the tracking system is guaranteed under the set of feedback control laws $u_i(k,t)=K_i(k)X(t)$, $i\in\Im$, then the cost function sequence $V_k(X,i)=X^T(t)\bar{P}_i(k)X(t)$ is finite as $0\le V_k(X,i)\le B_v$ for any $k=0,\ldots,+\infty$ in the iterations, and B_v is a positive constant.

Define $\tilde{V}_k(X,i)=V_{k+1}(X,i)-V_k(X,i)$, which can be rewritten by $\tilde{V}_k(X,i)=X^{\mathrm{T}}(t)\bar{P}_i(k+1)X(t)-X^{\mathrm{T}}(t)\bar{P}_i(k)X(t)=X^{\mathrm{T}}(t)[\bar{P}_i(k+1)-\bar{P}_i(k)]X(t)$. Hence, one has $\tilde{V}_0(X,i)=V_1(X,i)-V_0(X,i)=V_1(X,i)\ge 0$ as $V_0(X,i)=0$.

By using (33) and weak infinitesimal operator $\mathfrak{L}$, it follows that

$$\begin{aligned}&\mathfrak{L}\tilde{V}_k(X,i)\\&\triangleq\lim_{\triangle\to 0}\frac{1}{\triangle}[E\{\tilde{V}_k(X(t+\triangle),r(t+\triangle))|X(t),r(t)=i\}\\&\qquad-\tilde{V}_k(X(t),r(t)=i)]\\&=\lim_{\triangle\to 0}\frac{1}{\triangle}[E\{(V_{k+1}(e(t+\triangle),r(t+\triangle))\\&\qquad-V_k(e(t+\triangle),r(t+\triangle)))|e(t),r(t)=i\}\\&\qquad-(V_{k+1}(e(t),r(t)=i)-V_k(e(t),r(t)=i))]\\&=X^{\mathrm{T}}(t)\Bigg[(\bar{\mathcal{A}}_i+\bar{B}_i\bar{K}_i(k+1))^{\mathrm{T}}(\bar{P}_i(k+1)-\bar{P}_i(k))\\&\qquad+(\bar{P}_i(k+1)-\bar{P}_i(k))(\bar{\mathcal{A}}_i+\bar{B}_i\bar{K}_i(k+1))\\&\qquad+\sum_{j=1j\ne i}^{N}\pi_{ij}(\bar{P}_j(k+1)-\bar{P}_j(k))\Bigg]X(t)\\&=X^{\mathrm{T}}(t)\Bigg[(\bar{\mathcal{A}}_i+\bar{B}_i\bar{K}_i(k+1))^{\mathrm{T}}\bar{P}_i(k+1)\\&\qquad+\bar{P}_i(k+1)(\bar{\mathcal{A}}_i+\bar{B}_i\bar{K}_i(k+1))\\&\qquad+\sum_{j=1j\ne i}^{N}\pi_{ij}\bar{P}_j(k+1)-(\bar{\mathcal{A}}_i+\bar{B}_i\bar{K}_i(k+1))^{\mathrm{T}}\\&\qquad\times\bar{P}_i(k)-\bar{P}_i(k)(\bar{\mathcal{A}}_i+\bar{B}_i\bar{K}_i(k+1))\\&\qquad-\sum_{j=1j\ne i}^{N}\pi_{ij}\bar{P}_j(k)\Bigg]X(t)\\&=X^{\mathrm{T}}(t)\Bigg[-\bar{Q}_i-\lambda\ln\eta\bar{P}_i(k)-\bar{K}_i^{\mathrm{T}}(k+1)R_i\bar{K}_i(k+1)\\&\qquad-(\bar{\mathcal{A}}_i-\bar{B}_i\bar{K}_i(k))^{\mathrm{T}}\bar{P}_i(k)-\bar{P}_i(k)(\bar{\mathcal{A}}_i-\bar{B}_i\bar{K}_i(k))\\&\qquad-\sum_{j=1j\ne i}^{N}\pi_{ij}\bar{P}_j(k)-\bar{K}_i^{\mathrm{T}}(k)\bar{B}_i^{\mathrm{T}}\bar{P}_i(k)\\&\qquad-\bar{P}_i(k)\bar{B}_i\tilde{K}_i(k)\Bigg]X(t).\end{aligned}\tag{35}$$

Using (33) and mathematical induction, it becomes

$$\begin{aligned}&\mathfrak{L}\tilde{V}_k(X,i)\\&=X^{\mathrm{T}}(t)\big[-\bar{Q}_i-\lambda\ln\eta\bar{P}_i(k)-\bar{K}_i^{\mathrm{T}}(k+1)R_i\bar{K}_i(k+1)\\&\qquad+\bar{Q}_i+\lambda\ln\eta\bar{P}_i(k)+\bar{K}_i^{\mathrm{T}}(k)R_i\bar{K}_i(k)\\&\qquad+\tilde{K}_i^{\mathrm{T}}(k)R_i\tilde{K}_i(k)+\bar{K}_i^{\mathrm{T}}(k+1)R_i\bar{K}_i(k+1)\\&\qquad-\bar{K}_i^{\mathrm{T}}(k)R_i\bar{K}_i(k)\big]X(t)\\&=X^{\mathrm{T}}(t)\big[\tilde{K}_i^{\mathrm{T}}(k)R_i\tilde{K}_i(k)\big]X(t)\ge 0\end{aligned}\tag{36}$$

where $\tilde{K}_i(k)=\bar{K}_i(k+1)-\bar{K}_i(k)$.

Since Ω^* is a compact set and $\tilde{V}_k(X,i)=V_{k+1}(X,i)-V_k(X,i)\ge 0$ for any k in the iterations, the set $\{V_k|V_k(X,i)=X^{\mathrm{T}}(t)\bar{P}_i(k)X(t)\ge 0,k=1,\ldots\}$ is a bounded monotonic nondecreasing sequence; thus, by Dini's theorem [41], the sequence $V_k(X,i)$ will uniform pointwise converge to the optimal solution $V^*(X,i)$. Finally, the uniform convergence of feedback control sequence $u_i(k,t)$ is also achieved as the procedure.

It can be concluded that the quadratic Lyapunov equation $V_k(X,i)\to V^*(X,i)$ as $k\to\infty$, and the feedback control laws also gets converged as $u_i(k,t)\to u_i^*(t)$, which completes the proof. ■

The stability and convergence of the novel IRL-based tracking control algorithm are achieved by the theorems above, and the tracking control objective of the CMJS is obtained based on the developed online control policy, where the Markovian jump tracking errors are guaranteed to be stochastically stable. It means that all tracking errors converge to zero under the feedback control (30) designed in Algorithm 2.

Remark 5: Based on the IRL technique, the solution of CARE will be solved by the novel IRL parallel tracking control algorithm, which has effectively eliminated the crucial information requirements of stationary transition probability π_{ij} and the system matrix A_i in the traditional offline method. Both the stability and convergence of the designed optimal tracking control algorithm for CMJS are proved to be true, and the tracking objective is achieved as desired.

V. Simulation Results

In this section, two simulations are applied to verify the novel algorithms, where the continuous harmonic waveforms are utilized as desired references. According to the designed tracking control algorithms, the optimal solutions are obtained and the Markovian jump tracking dynamics get well controlled; then, the tracking errors converge to zero as the tracking objective is achieved.

A. Example 1: The Transition-Probability-Based Offline Algorithm for CMJS Tracking Problem

In this case, let us consider a fourth-order jump linear control problem [14], [31] to illustrate the feasibility and effectiveness of the proposed offline algorithm. The matrix coefficients for $N = 2$ of the CMJS (1) are given as

$$A_1 = \begin{bmatrix} -2.1051 & -1.1648 & 0.9347 & 0.5194 \\ -0.0807 & -2.8949 & 0.3835 & 0.8310 \\ 0.6914 & 10.5940 & -36.8199 & 3.8560 \\ 1.0692 & 13.4230 & 22.1185 & -13.1801 \end{bmatrix}$$

$$A_2 = \begin{bmatrix} -2.6430 & -1.2497 & 0.5269 & 0.6539 \\ -0.7910 & -2.8570 & 0.0920 & 0.4160 \\ 21.0357 & 22.8659 & -26.4655 & -1.7214 \\ 27.3096 & 7.8736 & -3.8604 & -29.5345 \end{bmatrix}$$

$$B_1 = \begin{bmatrix} 0.7564 \\ 0.9910 \\ 9.8255 \\ 7.2266 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 0.3653 \\ 0.2470 \\ 7.5336 \\ 6.5152 \end{bmatrix}, \quad \Pi = \begin{bmatrix} -0.6 & 0.6 \\ 0.4 & -0.4 \end{bmatrix}.$$

Thus, for a fourth-order Markov jump system, we select the continuous harmonic waveform as reference trajectory

$$\begin{cases} x_{d1} = 0.5\sin(\sqrt{5}t) \\ x_{d2} = 0.5\sqrt{5}\cos(\sqrt{5}t) \\ x_{d3} = 0.5\sin(\sqrt{5}t) \\ x_{d4} = 0.5\sqrt{5}\cos(\sqrt{5}t) \end{cases}$$

which can be further presented as

$$\dot{x}_d = Fx_d = \begin{bmatrix} \dot{x}_{d1} \\ \dot{x}_{d2} \\ \dot{x}_{d3} \\ \dot{x}_{d4} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -5 & 0 \end{bmatrix} \begin{bmatrix} x_{d1} \\ x_{d2} \\ x_{d3} \\ x_{d4} \end{bmatrix}. \quad (37)$$

By using Algorithm 1, the steady-state control input (16a) is derived from (12) as

$$u_i^c(t) = B_i^+[F - A_i]x_d(t). \quad (38)$$

Then, we select the symmetric positive definite matrices $Q_1 = \begin{bmatrix} 5 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 5 \end{bmatrix}$ and $Q_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$, $R_1 = R_2 = 1$ for the corresponding tracking cost function (13), by applying the transition-probability-based offline iteration process,

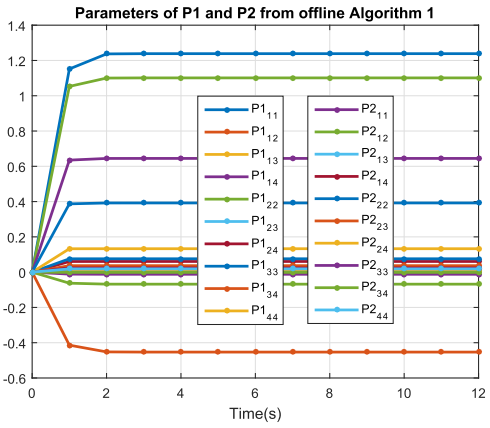

Fig. 1. Parameters of $\mathcal{P}_1$ and $\mathcal{P}_2$ from the designed offline tracking algorithm.

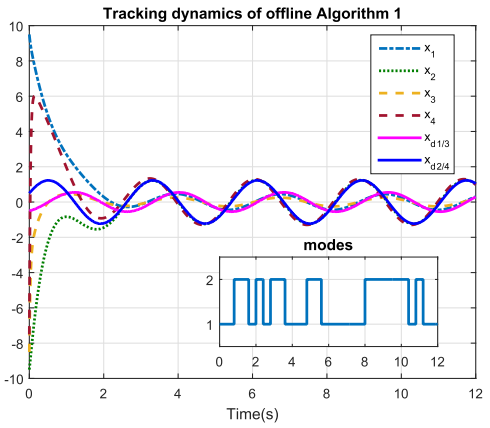

Fig. 2. Tracking trajectories from the designed offline tracking algorithm.

the CARE matrix solutions of are solved as

$$\mathcal{P}_1 = \begin{bmatrix} 1.2386 & -0.4532 & 0.0081 & -0.0121 \\ -0.4532 & 1.1012 & 0.0279 & 0.0606 \\ 0.0081 & 0.0279 & 0.0758 & 0.0367 \\ -0.0121 & 0.0606 & 0.0367 & 0.1328 \end{bmatrix}$$

$$\mathcal{P}_2 = \begin{bmatrix} 0.6447 & -0.0681 & 0.0184 & 0.0237 \\ -0.0681 & 0.3923 & 0.0128 & 0.0055 \\ 0.0184 & 0.0128 & 0.0197 & -0.0014 \\ 0.0237 & 0.0055 & -0.0014 & 0.0183 \end{bmatrix}$$

and the optimal feedback control gains are

$$\mathcal{K}_1 = \begin{bmatrix} 0.4805 & 1.4603 & 1.0441 & 1.3716 \end{bmatrix}^{\mathrm{T}}$$
$$\mathcal{K}_2 = \begin{bmatrix} 0.5114 & 0.2045 & 0.1490 & 0.1187 \end{bmatrix}^{\mathrm{T}}.$$

The tuning process of the matrix parameters $\mathcal{P}_1$ and $\mathcal{P}_2$ for CARE is shown in Fig. 1, where the convergence can be found clearly during the iteration learning process. Thus, it demonstrates that the offline Algorithm 1 obtains the optimal control solution.

By using the obtained control policy from the developed Algorithm 1, which consists of control inputs (16a) and (16b), the system tracking trajectory is presented in Fig. 2, where the references are well tracked as desired.

Note that Algorithm 1 is the transition-probability-based offline method, which has been pointed out earlier, and it

is designed by using the information of system matrices and transition probabilities to solve the CARE for optimal tracking control of the Markov jump systems. For the iteration learning process, the exact stationary transition probability crucial and the requirement of system information are hardly measured or obtained accurately in practical applications, which is a shortcoming in the offline computational iterations. Then, the designed Algorithm 2 will reduce these requirements of system information in the homogeneous Markovian process via the integral RL method.

B. Example 2: The Transition-Probability-Free IRL Algorithm for CMJS Tracking Problem

In the case, consider the modified mass, spring, and damper dynamic [19] in a large-scale device, which is applied and presented as

$$\begin{aligned}\dot{x}_1 &= x_2\\ \dot{x}_2 &= -\frac{k(r(t))}{m}x_1 - \frac{c(r(t))}{m}x_2 + \frac{1}{m}u\end{aligned} \tag{39}$$

where x_1 is the position, x_2 is the velocity, m is the mass of the object, and the parameters $k(r(t))$ and $c(r(t))$ denote the mode-based stiffness of this spring and the damping. The values of the parameters are given by $k = 5$ (N · m) and $c = 0.05$ when $r = 1$, $k = 6$ (N · m) and $c = 0.01$ when $r = 2$, and $m = 5$ (kg).

Select the reference trajectory also as the continuous harmonic waveform by the form

$$\begin{cases} x_{d1} = 0.5\sin(\sqrt{5}t) \\ x_{d2} = 0.5\sqrt{5}\cos(\sqrt{5}t) \end{cases} \tag{40}$$

and suppose the stiffness parameter and damping in the system have different modes within a homogeneous Markovian process, we augment the new tracking system as (23), and then it becomes

$$\dot{X}(t) = \begin{bmatrix} A_i & A_i - F \\ \mathbf{0} & F \end{bmatrix} X(t) + \begin{bmatrix} B_i \\ \mathbf{0} \end{bmatrix} u_i(t) \equiv \bar{A}_i X + \bar{B}_i u_i \tag{41}$$

where $A_1 = \begin{bmatrix} 0 & 1 \\ -5 & -0.05 \end{bmatrix}$, $A_2 = \begin{bmatrix} 0 & 1 \\ -6 & -0.01 \end{bmatrix}$, $F = \begin{bmatrix} 0 & 1 \\ -5 & 0 \end{bmatrix}$, $B_1 = B_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$, $N = 2$, and $\Pi = \begin{bmatrix} -0.6 & 0.6 \\ 0.4 & -0.4 \end{bmatrix}$ is the stationary transition probability matrix.

It should be pointed out that based on the designed Algorithm 2, the system matrices A_i, $i \in \Im$, and F and the stationary transition probability π_{ij} are not used during the solving process. Then, we select the symmetric positive definite matrices $\bar{Q}_1 = \begin{bmatrix} 10 & 0 & 0 & 0 \\ 0 & 10 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ and $\bar{Q}_2 = \begin{bmatrix} 5 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 \\ 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0.5 \end{bmatrix}$, $R_1 = R_2 = 1$, $\eta = 0.5$, $\lambda = 0.1$, $r(0) = 1$ for the discounted cost function (24), and initial states $X(0) = [1, -1, -0.11, 0.13]^{\mathrm{T}}$.

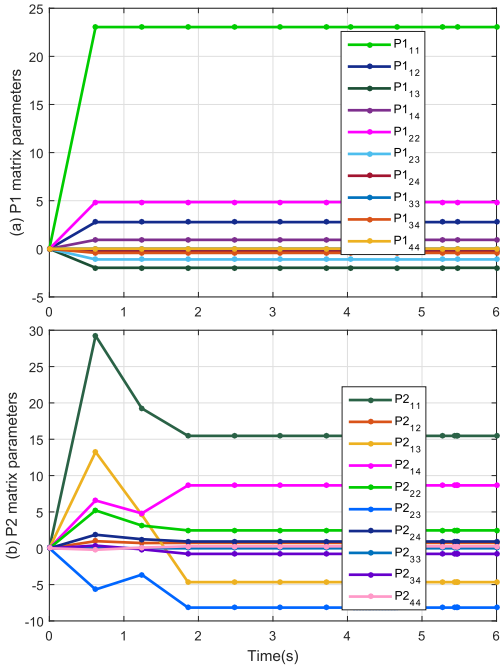

Fig. 3. Evolution of the parameters of matrices $\bar{\mathcal{P}}_1$ and $\bar{\mathcal{P}}_2$ from the proposed IRL parallel tracking control algorithm. Updating procedure of (a) $\bar{\mathcal{P}}_1$ and (b) $\bar{\mathcal{P}}_2$.

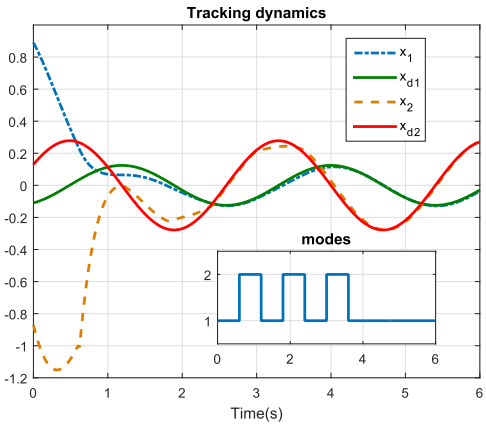

Fig. 4. Trajectories of the tracking states from the novel IRL algorithm.

The updating procedures of the positive definite matrix parameters can be seen clearly in Fig. 3, where the parameters of matrices $\bar{\mathcal{P}}_1$ and $\bar{\mathcal{P}}_2$ get converged in the online learning process. Besides, both the stability and convergence of the proposed algorithm are achieved, and the optimal control policy (30) is obtained without using the system matrices A_i, $i \in \Im$, and F and the stationary transition probability π_{ij}. Along with the obtained solutions of CARE from designed Algorithm 2, the optimal feedback control gains for the Markov jump tracking system are achieved as $\mathcal{K}_1 = [1.3874, 4.8540, -0.5516, -0.1280]$ and $\mathcal{K}_2 = [0.3709, 2.4585, -4.0824, 0.4636]$.

To further explain the effectiveness of the designed transition-probability-free IRL algorithm 2, the evolution of tracking states in the new augmented CMJS is presented in Fig. 4, where the desired reference trajectories get well

tracked under the proposed IRL control scheme. Moreover, the Markovian jump modes are also displayed in Fig. 4, and it is obvious that along with the propagating time, the solutions of the augmented tracking control problem for CMJS are achieved successfully as we desired.

VI. Conclusion

In this article, a novel transition-probability-free optimal tracking control algorithm has been proposed for the continuous-time Markov jump system via the IRL technique. By adopting the exact information of the system and transition probability, an offline tracking control algorithm has been presented, where the control policy consists of steady-state control input and a feedback control input. To overcome the shortcomings and crucial requirements of system information and stationary transition probability in the homogeneous Markovian process during the offline iteration process, the transition-probability-free IRL optimal tracking control algorithm has been proposed successfully. For the designed algorithms, both the convergence of the CARE and the stochastic stability of the tracking error dynamic have been proved and achieved by theorems. In addition, we have verified the offline parallel tracking and novel transition-probability-free IRL tracking control algorithms in applications. Then, the simulation results have demonstrated the effectiveness and applicability of the developed algorithms.

References

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

[2] D. Silver *et al.*, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016.

[3] F.-Y. Wang, H. Zhang, and D. Liu, "Adaptive dynamic programming: An introduction," *IEEE Comput. Intell. Mag.*, vol. 4, no. 2, pp. 39–47, May 2009.

[4] P. A. Ioannou, H. Xu, and B. Fidan, "Identification and high bandwidth control of hard disk drive servo systems based on sampled data measurements," *IEEE Trans. Control Syst. Technol.*, vol. 15, no. 6, pp. 1089–1095, Nov. 2007.

[5] S. Vrkalovic, E.-C. Lunca, and I.-D. Borlea, "Model-free sliding mode and fuzzy controllers for reverse osmosis desalination plants," *Int. J. Artif. Intell.*, vol. 16, no. 2, pp. 208–222, 2018.

[6] D. Liu, H. Javaherian, O. Kovalenko, and T. Huang, "Adaptive critic learning techniques for engine torque and air–fuel ratio control," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 38, no. 4, pp. 988–993, Aug. 2008.

[7] H. Modares, F. L. Lewis, and M.-B. Naghibi-Sistani, "Integral reinforcement learning and experience replay for adaptive optimal control of partially-unknown constrained-input continuous-time systems," *Automatica*, vol. 50, no. 1, pp. 193–202, Jan. 2014.

[8] T. Çimen and S. P. Banks, "Nonlinear optimal tracking control with application to super-tankers for autopilot design," *Automatica*, vol. 40, no. 11, pp. 1845–1863, Nov. 2004.

[9] H. Zhang, L. Cui, X. Zhang, and Y. Luo, "Data-driven robust approximate optimal tracking control for unknown general nonlinear systems using adaptive dynamic programming method," *IEEE Trans. Neural Netw.*, vol. 22, no. 12, pp. 2226–2236, Dec. 2011.

[10] Z. Ni, H. He, and J. Wen, "Adaptive learning in tracking control based on the dual critic network design," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 6, pp. 913–928, Jun. 2013.

[11] H.-N. Wu and B. Luo, "Simultaneous policy update algorithms for learning the solution of linear continuous-time H_∞ state feedback control," *Inf. Sci.*, vol. 222, pp. 472–485, Feb. 2013.

[12] Y. Zhu, D. Zhao, H. He, and J. Ji, "Event-triggered optimal control for partially unknown constrained-input systems via adaptive dynamic programming," *IEEE Trans. Ind. Electron.*, vol. 64, no. 5, pp. 4101–4109, May 2017.

[13] D. L. Kleinman, "On an iterative technique for Riccati equation computations," *IEEE Trans. Autom. Control*, vol. AC-13, no. 1, pp. 114–115, Feb. 1968.

[14] Z. Gajic and I. Borno, "Lyapunov iterations for optimal control of jump linear systems at steady state," *IEEE Trans. Autom. Control*, vol. 40, no. 11, pp. 1971–1975, Nov. 1995.

[15] S. Preitl, R.-E. Precup, Z. Preitl, S. Vaivoda, S. Kilyeni, and J. K. Tar, "Iterative feedback and learning control. Servo systems applications," *IFAC Proc. Volumes*, vol. 40, no. 8, pp. 16–27, Jul. 2007.

[16] K. C. C. Chan, V. Lee, and H. Leung, "Generating fuzzy rules for target tracking using a steady-state genetic algorithm," *IEEE Trans. Evol. Comput.*, vol. 1, no. 3, pp. 189–200, Sep. 1997.

[17] H. Zhang, Q. Wei, and 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. B, Cybern.*, vol. 38, no. 4, pp. 937–942, Aug. 2008.

[18] Y.-M. Park, M.-S. Choi, and K. Y. Lee, "An optimal tracking neuro-controller for nonlinear dynamic systems," *IEEE Trans. Neural Netw.*, vol. 7, no. 5, pp. 1009–1110, Sep. 1996.

[19] H. Modares and F. L. Lewis, "Optimal tracking control of nonlinear partially-unknown constrained-input systems using integral reinforcement learning," *Automatica*, vol. 50, no. 7, pp. 1780–1792, Jul. 2014.

[20] B. Luo, D. Liu, T. Huang, and D. Wang, "Model-free optimal tracking control via critic-only Q-learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 10, pp. 2134–2144, Oct. 2016.

[21] Z.-G. Wu, Y. Shen, P. Shi, Z. Shu, and H. Su, "H_∞ control for 2-D Markov jump systems in Roesser model," *IEEE Trans. Autom. Control*, vol. 64, no. 1, pp. 427–432, Jan. 2019.

[22] J. Dong and G.-H. Yang, "Robust H_2 control of continuous-time Markov jump linear systems," *Automatica*, vol. 44, no. 5, pp. 1431–1436, May 2008.

[23] T. Bian and Z.-P. Jiang, "A tool for the global stabilization of stochastic nonlinear systems," *IEEE Trans. Autom. Control*, vol. 62, no. 4, pp. 1946–1951, Apr. 2017.

[24] H.-N. Wu and K.-Y. Cai, "Mode-independent robust stabilization for uncertain Markovian jump nonlinear systems via fuzzy control," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 36, no. 3, pp. 509–519, Jun. 2005.

[25] P. Shi, E. K. Boukas, and R. K. Agarwal, "Control of Markovian jump discrete-time systems with norm bounded uncertainty and unknown delay," *IEEE Trans. Autom. Control*, vol. 44, no. 11, pp. 2139–2144, Nov. 1999.

[26] J. Song, S. He, F. Liu, Y. Niu, and Z. Ding, "Data-driven policy iteration algorithm for optimal control of continuous-time Itô stochastic systems with Markovian jumps," *IET Control Theory Appl.*, vol. 10, no. 12, pp. 1431–1439, Aug. 2016.

[27] Y. Ji and H. J. Chizeck, "Controllability, stabilizability, and continuous-time Markovian jump linear quadratic control," *IEEE Trans. Autom. Control*, vol. 35, no. 7, pp. 777–788, Jul. 1990.

[28] G. Nakura, "Stochastic optimal tracking with preview by state feedback for linear discrete-time Markovian jump systems," *Int. J. Innovative Comput. Inf. Control*, vol. 6, no. 1, pp. 15–28, Jan. 2010.

[29] K. D. Running and N. C. Martins, "Optimal preview control of Markovian jump linear systems," *IEEE Trans. Autom. Control*, vol. 54, no. 9, pp. 2260–2266, Sep. 2009.

[30] F.-Y. Wang, Y. Yuan, C. Rong, and J. J. Zhang, "Parallel blockchain: An architecture for CPSS-based smart societies," *IEEE Trans. Comput. Social Syst.*, vol. 5, no. 2, pp. 303–310, Jun. 2018.

[31] I. G. Ivanov, "On some iterations for optimal control of jump linear equations," *Nonlinear Anal., Theory, Methods Appl.*, vol. 69, no. 11, pp. 4012–4024, 2008.

[32] I. Borno, "Parallel computation of the solutions of coupled algebraic Lyapunov equations," *Automatica*, vol. 31, no. 9, pp. 1345–1347, Sep. 1995.

[33] O. L. V. Costa, J. B. R. do Val, and J. C. Geromel, "Continuous-time state-feedback H_2-control of Markovian jump linear systems via convex analysis," *Automatica*, vol. 35, no. 2, pp. 259–268, Feb. 1999.

[34] Z. Li, B. Zhou, J. Lam, and Y. Wang, "Positive operator based iterative algorithms for solving Lyapunov equations for Itô stochastic systems with Markovian jumps," *Appl. Math. Comput.*, vol. 217, no. 21, pp. 8179–8195, Jul. 2011.

[35] J. Song *et al.*, "A new iterative algorithm for solving H_∞ control problem of continuous-time Markovian jumping linear systems based on online implementation," *Int. J. Robust Nonlinear Control*, vol. 26, no. 17, pp. 3737–3754, Nov. 2016.

[36] X. Niu, X. Gao, and Y. Weng, "Data-driven sliding mode tracking control for unknown Markovian jump non-linear systems," *IET Control Theory Appl.*, vol. 11, no. 16, pp. 2716–2723, Nov. 2017.
[37] R. Sakthivel, S. Harshavarthini, R. Kavikumar, and Y.-K. Ma, "Robust tracking control for fuzzy Markovian jump systems with time-varying delay and disturbances," *IEEE Access*, vol. 6, pp. 66861–66869, 2018.
[38] X. Zhong, H. He, H. Zhang, and Z. Wang, "Optimal control for unknown discrete-time nonlinear Markov jump systems using adaptive dynamic programming," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 12, pp. 2141–2155, Dec. 2014.
[39] T. Bian, Y. Jiang, and Z.-P. Jiang, "Adaptive dynamic programming for stochastic systems with state and control dependent noise," *IEEE Trans. Autom. Control*, vol. 61, no. 12, pp. 4170–4175, Dec. 2016.
[40] I. K. Argyros, Y. J. Cho, and S. Hilout, *Numerical Methods for Equations and Its Applications*. Boca Raton, FL, USA: CRC Press, 2012.
[41] R. G. Bartle and D. R. Sherbert, *Introduction to Real Analysis*, 3rd ed. New York, NY, USA: Wiley, 2000.

Kun Zhang (S'18) received the B.S. degree in mathematics and applied mathematics from Hebei Normal University, Shijiazhuang, China, in 2012. He is currently pursuing the Ph.D. degree in control theory and control engineering with Northeastern University, Shenyang, China.

His main research interests include nonlinear systems, reinforcement learning, intelligent control, adaptive dynamic programming, optimization algorithm, and their industrial applications.

Hua-guang Zhang (M'03–SM'04–F'14) received the B.S. and M.S. degrees in control engineering from the Northeast Dianli University, Jilin, China, in 1982 and 1985, respectively, and the Ph.D. degree in thermal power engineering and automation from Southeast University, Nanjing, China, in 1991.

He has authored or coauthored over 280 journal and conference articles, six monographs, and coinvented 90 patents. His current research interests include fuzzy control, stochastic system control, neural network-based control, nonlinear control, and their applications.

Dr. Zhang was a recipient of the Outstanding Youth Science Foundation Award from the National Natural Science Foundation Committee of China in 2003 and the IEEE TRANSACTIONS ON NEURAL NETWORKS 2012 Outstanding Paper Award. He was named the Cheung Kong Scholar by the Education Ministry of China in 2005. He is also the E-Letter Chair of the IEEE CIS Society and the Former Chair of the Adaptive Dynamic Programming and Reinforcement Learning Technical Committee of the IEEE Computational Intelligence Society. He was an Associate Editor of the IEEE TRANSACTIONS ON FUZZY SYSTEMS from 2008 to 2013. He is also an Associate Editor of *Automatica*, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON CYBERNETICS, and *Neurocomputing*.

Yuliang Cai received the B.S. degree in information and computing science from Ludong University, Yantai, China, in 2014, and the M.S. degree in control theory and control engineering from the Dalian University of Technology, Dalian, China, in 2017. She is currently pursuing the Ph.D. degree in control theory and control engineering with Northeastern University, Shenyang, China.

Her research interests include multiagent system control, fuzzy control and adaptive dynamic programming.

Rong Su (M'11–SM'14) received the B.E. degree from the University of Science and Technology of China, Hefei, China, in 1997, and the Master of Applied Science and Ph.D. degree from the University of Toronto, Toronto, ON, USA, in 2000 and 2004, respectively.

He was with the University of Waterloo, Waterloo, ON, Canada, and the Eindhoven University of Technology, Eindhoven, The Netherlands. He joined Nanyang Technological University, Singapore, in 2010, where he is currently an Associate Professor with the School of Electrical and Electronic Engineering. His research interests include multiagent systems, discrete-event system theory, model-based fault diagnosis, control and optimization for complex networked systems with applications in flexible manufacturing, intelligent transportation, human–robot interface, power management, and green building. In these areas, he has more than 180 journal and conference publications, two granted patents, and four patent applications.

Dr. Su is also the Chair of the Technical Committee on Smart Cities of the IEEE Control Systems Society. He is also an Associate Editor of *Automatica*, the *Journal of Discrete Event Dynamic Systems: Theory and Applications*, and the *Journal of Control and Decision*.