SCIENCE CHINA
Information Sciences

• LETTER •

November 2024, Vol. 67, Iss. 11, 219203:1–219203:2
https://doi.org/10.1007/s11432-024-4139-6

Policy iteration-based adaptive optimal control for Markov jump systems: a transition-probability-free asynchronous approach

Weidi CHENG[1,2], Chengcheng REN[1,2], Shuping HE[1,2,3*] & Changyin SUN[4]

[1]*Key Laboratory of Intelligent Computing and Signal Processing (Ministry of Education),*
School of Electrical Engineering and Automation, Anhui University, Hefei 230601, China;
[2]*Anhui Engineering Laboratory of Human-Robot Integration System and Intelligent Equipment,*
School of Electrical Engineering and Automation, Anhui University, Hefei 230601, China;
[3]*Information Materials and Intelligent Sensing Laboratory of Anhui Province, Hefei 230601, China;*
[4]*Engineering Research Center of Autonomous Unmanned System Technology (Ministry of Education),*
School of Artificial Intelligence, Anhui University, Hefei 230601, China

Received 25 March 2024/Revised 21 June 2024/Accepted 26 August 2024/Published online 18 October 2024

The controller and filter design problems of Markov jump systems (MJSs) have gained significant attention over the past few decades. These studies include various aspects, including stochastic stabilization [1], optimal tracking control [2], and dissipative filter design [3]. Although numerous publications address the optimal controller design for MJSs, the issue of hidden MJSs, particularly those with mismatched jumping modes between the system and the controller, has rarely been explored.

In this study, we propose a policy iteration-based adaptive optimal control scheme for MJSs using a transition-probability-free asynchronous approach. The main contributions are as follows: First, an asynchronous infinite horizon performance index is established, the weight matrix and control policy jumping are based on the detected mode under conditional probability. Second, an asynchronous policy iteration technique is introduced to iteratively solve the coupled algebraic Riccati equations (CAREs) using online measured state and input information without the need for system matrices and coupled transition probability information.

Preliminaries. Consider a probability space $(\Omega, \Upsilon, \mathrm{H})$. We examine a class of continuous-time MJSs described by

$$\dot{x} = A_{\rho(t)}x + B_{\rho(t)}u, \tag{1}$$

where $x \in \mathbb{R}^n$ is the system state, $u \in \mathbb{R}^m$ is the control input. $A_{\rho(t)} \in \mathbb{R}^{n\times n}$ and $B_{\rho(t)} \in \mathbb{R}^{n\times m}$ are unknown constant matrices. Here, $\{\rho(t)\}$ denotes the system mode, characterized by a homogeneous Markovian process with right-continuous trajectories in a finite discrete state space $\Xi = \{1, 2, \ldots, \mathcal{N}\}$ with a transition probability matrix $\Pi = \{\phi_{ij}\}_{\mathcal{N}\times\mathcal{N}}$. The probabilities Π are defined by

$$\mathbb{P}\{\rho(t+\iota) = j|\rho(t) = i\} = \begin{cases} \phi_{ij}\iota + \Delta(\iota), & i \neq j, \\ 1 + \phi_{ii}\iota + \Delta(\iota), & i = j, \end{cases} \tag{2}$$

where $\iota > 0$, $\lim_{\iota\to 0}\Delta(\iota)\iota^{-1} = 0$, and $\phi_{ij} \geqslant 0$ is the transition rate from mode i to j when $i \neq j$; $i, j \in \Xi$, and $\phi_{ii} = -\sum_{j=1, j\neq i}^{\mathcal{N}} \phi_{ij}$.

Considering potential inaccuracies in detecting the system mode $\rho(t)$, a hidden Markov model is employed to describe the connection between the detected mode and the actual system mode. The detected mode is denoted by $\eta(t)$, with values in $\Lambda = \{1, 2, \ldots, \mathcal{M}\}$. It is ensured that $\Lambda \subseteq \Xi$ and $\Lambda \neq \emptyset$. The conditional probability matrix Θ defines the relationship between the detected mode and the system mode as $\mathbb{P}\{\eta(t) = l|\rho(t) = i\} = \tau_{il}$ in which $\tau_{il} \in [0, 1]$, $\sum_{l=1}^{\mathcal{M}} \tau_{il} = 1$, $\forall l \in \Lambda$ and $i \in \Xi$.

To achieve asynchronous optimal control for MJSs (1), the objective is to identify a mode-independent control policy based on the detected mode $\eta(t)$ rather than the actual mode $\rho(t)$. Consequently, $u_{\eta(t)}$ can be expressed as $u_{\eta(t)} \triangleq -\xi_{\eta(t)}(t, x) = -K_{\eta(t)}x$, where $\xi_{\eta(t)} : [0, \infty) \times \mathbb{R}^n \times \Lambda \to \mathbb{R}^m$ represents the admissible control policy. The asynchronous infinite horizon performance index related MJSs (1) is then defined as $\mathcal{V}_{\eta(t)}(t, x) = \mathbb{E}\{\int_0^\infty [x^{\mathrm{T}}Q_{\eta(t)}x + u_{\eta(t)}^{\mathrm{T}}R_{\eta(t)}u_{\eta(t)}]\mathrm{d}t|x_0, \eta_0\}$ with $Q_{\eta(t)} \succ 0$ and $R_{\eta(t)} \succ 0$.

Subsequently, the infinite horizon asynchronous optimal control problem is reduced to finding an admissible control policy, denoted as $u_{\eta(t)}^* \triangleq \underset{u_{\eta(t)}, t_0\leqslant t\leqslant\infty}{\operatorname{argmin}} \mathcal{V}_{\eta(t)}(t_0, x_{t_0}, u_{\eta(t)})$.

For simplicity, when $\rho(t) = i$, $i \in \Xi$, let $A_{\rho(t)}$ denote A_i. Similarly, when $\eta(t) = l$, $l \in \Lambda$, $K_{\eta(t)}$ is represented as K_l. These notations apply to other functions as well.

Lemma 1 ([4]). Consider any initial stabilizing gain matrix $K_0 \in \mathbb{R}^{m\times n}$, $\beta \in \mathbb{Z}_+$ and let $P^{(\beta)}$ denote the symmetric positive definite solution of the following Lyapunov equation $[A - BK^{(\beta)}]^{\mathrm{T}}P^{(\beta)} + P^{(\beta)}[A - BK^{(\beta)}] + (K^{(\beta)})^{\mathrm{T}}RK^{(\beta)} + Q = 0$ with $K^{(\beta)} = R^{-1}B^{\mathrm{T}}P^{(\beta-1)}$. Then, the following statements hold: (1) $A - BK^{(\beta)}$ is Hurwitz; (2) $P^* \leqslant P^{(\beta+1)} \leqslant P^{(\beta)}$; (3) $\lim_{\beta\to\infty}K^{(\beta)} = K^*$, $\lim_{\beta\to\infty}P^{(\beta)} = P^*$.

Remark 1. It is evident from $u_l = -K_lx = -R_l^{-1}B_i^{\mathrm{T}}P_lx$ that precise knowledge of the original jump system matrix

[*] Corresponding author (email: shuping.he@ahu.edu.cn)

B_i remains necessary for the iterations. Therefore, the main challenge addressed in this study is proposing a data-driven learning strategy to achieve an asynchronous optimal control policy without relying on the original jump system mode matrices A_i and B_i.

Assumption 1. The precise values of $\{\rho(t)\}$ and $\{\eta(t)\}$ are known. The expectation in $\mathcal{V}_{\eta(t)}(t,x)$ is computed over the joint process $\{x,\rho(t),\eta(t)\}$, confirming the stochastic controllability of the MJSs described by (1).

Assumption 2. The system $(A_i, B_i, \sqrt{Q_l})$, $i \in \Xi$, $l \in \Lambda$ is stochastically detectable.

Algorithm implementation. We reconstruct the MJSs describe by (1) as follows:

$$\dot{x} = \tilde{A}_i^{(\beta)} x + B_i(K_l^{(\beta)} x + u_l), \tag{3}$$

where $\tilde{A}_i^{(\beta)} = A_i - B_i K_l^{(\beta)}$.

Let $K_l^{(0)}$ be a known initial stabilizing gain for (3). The corresponding asynchronous infinite horizon performance index is designed as $\mathcal{V}_l(t,x) = \mathbb{E}\{\int_t^\infty \mu^{\gamma(\tau-t)}(x^{\mathrm{T}} Q_l x + u_l^{\mathrm{T}} R_l u_l)\mathrm{d}\tau | x_t, \eta_t\}$, where $\mu \in (0,1)$ represents the discount factor, γ is a positive parameter, $R_l \succ 0$, $Q_l \succ 0$ are the weight matrices. Consequently, the relevant CAREs with $u_l = -K_l^{(\beta)} x$ yields $[\tilde{A}_i^{(\beta)} - B_i K_l^{(\beta)}]^{\mathrm{T}} P_l^{(\beta)} + P_l^{(\beta)}[\tilde{A}_i^{(\beta)} - B_i K_l^{(\beta)}] = -(K_l^{(\beta)})^{\mathrm{T}} R_l K_l^{(\beta)} - \sum_{j=1}^{\mathcal{N}} \sum_{l=1}^{\mathcal{M}} \phi_{ij} \tau_{il} P_j - \gamma \ln \mu P_l^{(\beta-1)} - Q_l$. For each iteration $\beta \in \mathbb{Z}_+$ and detected mode $l \in \Lambda$, we seek a sequence of symmetric positive definite matrices $P_l^{(\beta)}$ that satisfies the CAREs. We also derive a feedback gain matrix $K_l^{(\beta+1)} \in \mathbb{R}^{m\times n}$ using $K_l^{(\beta+1)} = R_l^{-1} B_i^{\mathrm{T}} P_l^{(\beta)}$. Along the solution of (3) by Lemma 1, we consider the initial stabilizing control signal $u_l^{(0)} = -K_l^{(0)} x + \theta$ where θ denotes the exploration noise [5]. The online implementation of the policy iteration approach is achieved as follows:

$$\begin{aligned} & x(t+o(t))^{\mathrm{T}} P_l^{(\beta)} x(t+o(t)) - x(t)^{\mathrm{T}} P_l^{(\beta)} x(t) \\ &= -\int_t^{t+o(t)} x^{\mathrm{T}} \bar{\mathcal{Q}}_l^{(\beta)} x \mathrm{d}\tau \\ &\quad + 2\int_t^{t+o(t)} (u_l + K_l^{(\beta)} x)^{\mathrm{T}} R_l K_l^{(\beta+1)} x \mathrm{d}\tau, \end{aligned} \tag{4}$$

where $\bar{\mathcal{Q}}_l^{(\beta)} = Q_l + K_l^{(\beta)\mathrm{T}} R_l K_l^{(\beta)} + \gamma \ln \mu P_l^{(\beta-1)}$.

Using the Kronecker product representation, we derive the following transition-probability-free adaptive asynchronous optimal control as Algorithm 1.

Algorithm 1 Adaptive asynchronous optimal control algorithm via policy iteration

```
Set the iteration step $\beta = 0$, the convergence error $\alpha > 0$, the initial control input $u_l^{(0)} = -K_l^{(0)} x + \theta$.
Compute $\sigma_{xx}$, $\xi_{xx}$ and $\xi_{xu}$ until $\mathrm{rank}([\xi_{xx}, \xi_{xu}]) = \frac{n(n+1)}{2} + mn$ holds.
repeat
    For $l \in \Lambda$ do
    Update $\Phi_l^{(\beta)}$ and $\Psi_l^{(\beta)}$ based on $\Phi_l^{(\beta)} = [\sigma_{xx}, -2\xi_{xx}(I_n \otimes K_l^{(\beta)\mathrm{T}} R_l) - 2\xi_{xu}(I_n \otimes R_l)]$, $\Psi_l^{(\beta)} = -\xi_{xx}\mathrm{vec}(\bar{\mathcal{Q}}_l^{(\beta)})$.
    Solve $P_l^{(\beta)}$ and $K_l^{(\beta+1)}$ from
        $$\begin{bmatrix} \bar{\Gamma}_l^{(\beta)} \\ \mathrm{vec}(K_l^{(\beta+1)}) \end{bmatrix} = (\Phi_l^{(\beta)\mathrm{T}} \Phi_l^{(\beta)})^{-1} \Phi_l^{(\beta)\mathrm{T}} \Psi_l^{(\beta)}. \tag{5}$$
    $\beta \leftarrow \beta + 1$.
until $\| P_l^{(\beta)} - P_l^{(\beta-1)} \| \leqslant \alpha$.
Update $u_l = -K_l^{(\beta)} x$ as the asynchronous optimal policy.
```

Remark 2. In Algorithm 1, $x \in \mathbb{R}^n \to \bar{\chi} \in \mathbb{R}^{\frac{1}{2}n(n+1)}$, $P_l \in \mathbb{R}^{n\times n} \to \bar{\Gamma}_l \in \mathbb{R}^{\frac{1}{2}n(n+1)}$, $\Phi_l^{(\beta)} \in \mathbb{R}^{\varepsilon \times [\frac{1}{2}n(n+1)+mn]}$, $\Psi_l^{(\beta)} \in \mathbb{R}^{\varepsilon}$. For $0 \leqslant t_0 < t_1 < \cdots < t_\varepsilon$, we define $\sigma_{xx} = [(\bar{\chi}_{t_1} - \bar{\chi}_{t_0}), \ldots, (\bar{\chi}_{t_\varepsilon} - \bar{\chi}_{t_{\varepsilon-1}})]^{\mathrm{T}} \in \mathbb{R}^{\varepsilon \times \frac{1}{2}n(n+1)}$, $\xi_{xx} = [\int_{t_0}^{t_1} x \otimes x \mathrm{d}\tau, \ldots, \int_{t_{\varepsilon-1}}^{t_\varepsilon} x \otimes x \mathrm{d}\tau]^{\mathrm{T}} \in \mathbb{R}^{\varepsilon \times n^2}$, $\xi_{xu} = [\int_{t_0}^{t_1} x \otimes u_l \mathrm{d}\tau, \ldots, \int_{t_{\varepsilon-1}}^{t_\varepsilon} x \otimes u_l \mathrm{d}\tau]^{\mathrm{T}} \in \mathbb{R}^{\varepsilon \times mn}$ where ε is a positive integer. More details can be found in Appendix C.

Theorem 1 (Convergence analysis)**.** For any detected mode $\eta(t) = l \in \Lambda$, under Assumptions 1 and 2, and starting from an initial stabilizing policy $K_l^{(0)}$, when the full column rank condition is satisfied, the sequence of solution pair $\{[P_l^{(\beta)}], [K_l^{(\beta)}]\}$ is uniquely determined by the equality in (5) and converges to the optimal pair (P_l^*, K_l^*).

Proof. For any detected mode $l \in \Lambda$, consider a stabilizing policy $K_l^{(\beta)}$; if $P_l^{(\beta)} = (P_l^{(\beta)})^{\mathrm{T}}$ is the solution of the CAREs, which satisfies Lemma 1, then $K_l^{(\beta+1)}$ is uniquely determined by $K_l^{(\beta+1)} = R_l^{-1} B_i^{\mathrm{T}} P_l^{(\beta)}$. Using equation (4), we obtain $P_l^{(\beta)}$ and $K_l^{(\beta+1)}$ that satisfy the equality in (5). Let $P_l = P_l^{\mathrm{T}} \in \mathbb{R}^{n\times n}$ and $K_l \in \mathbb{R}^{m\times n}$, so that $\Phi_l^{(\beta)} \begin{bmatrix} \bar{\Gamma}_l \\ \mathrm{vec}(K_l) \end{bmatrix} = \Psi_l^{(\beta)}$ holds. Then, $\bar{\Gamma}_l = \bar{\Gamma}_l^{(\beta)}$ and $\mathrm{vec}(K_l) = \mathrm{vec}(K_l^{(\beta+1)})$. Based on the full column rank condition, $P_l = P_l^{\mathrm{T}}$ and K_l are unique. Therefore, according to the definitions of $\bar{\Gamma}_l$ and $\mathrm{vec}(K_l)$, $P_l^{(\beta)} = P_l$ and $K_l^{(\beta+1)} = K_l$ are uniquely determined. Consequently, the convergence is established by Lemma 1.

Conclusion. A policy iteration-based adaptive optimal control algorithm using a transition-probability-free asynchronous approach has been developed for a class of MJSs. This algorithm approximates the optimal CARE solution without requiring prior knowledge of the system matrices. By employing a constructed discounted cost function, the coupled transition probabilities are no longer necessary. Finally, the convergence of the proposed algorithm is verified.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant No. 62073001), the Anhui Provincial Key Research and Development Project (Grant No. 2022i01020013), the University Synergy Innovation Program of Anhui Province (Grant No. GXXT-2021-010), and the Anhui Province Graduate Education Quality Project (Grant No. 2023xscx027).

Supporting information Appendixes A–D. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Wang G, Song S, Huang C. Stochastic stabilization of Markovian jump systems closed by a communication network: an auxiliary system approach. Sci China Inf Sci, 2023, 66: 202202

2 Zhang K, Zhang H G, Cai Y, et al. Parallel optimal tracking control schemes for mode-dependent control of coupled Markov jump systems via integral RL method. IEEE Trans Autom Sci Eng, 2020, 17: 1332–1342

3 Zhang X, He S, Stojanovic V, et al. Finite-time asynchronous dissipative filtering of conic-type nonlinear Markov jump systems. Sci China Inf Sci, 2021, 64: 152206

4 Kleinman D. On an iterative technique for Riccati equation computations. IEEE Trans Automat Contr, 1968, 13: 114–115

5 Jiang Y, Jiang Z P. Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics. Automatica, 2012, 48: 2699–2704