自动化学报
ACTA AUTOMATICA SINICA

基于无模型策略梯度强化学习的未知随机系统最优控制

杜城龙 韩洁 李繁飙 桂卫华

Model-free Policy Gradient-based Reinforcement Learning Algorithms for Optimal Control of Unknown Stochastic Systems

DU Cheng-Long, HAN Jie, LI Fan-Biao, GUI Wei-Hua

在线阅读 View online: https://doi.org/10.16383/j.aas.c250156

基于无模型策略梯度强化学习的未知随机系统最优控制

杜城龙 [1] 韩 洁 [1] 李繁飙 [1] 桂卫华 [1]

摘 要 针对一类未知动力学马尔科夫随机系统的最优控制问题, 提出两种无模型策略梯度强化学习算法. 首先, 针对模型信息部分未知的马尔科夫随机系统, 基于系统采样数据和耦合李雅普诺夫方程推导出无模型策略梯度的解析形式, 并提出一种部分无模型策略梯度强化学习最优控制算法, 实现对预设性能指标的直接最小化. 由于求解耦合李雅普诺夫方程和计算策略梯度的必要数据均可从系统采样数据同一轨迹提取, 而无需再额外收集采样数据, 降低了算法的采样复杂度. 进一步地, 为完全解除对马尔科夫随机系统模型信息的依赖, 通过随机摄动反馈增益估计策略梯度, 并提出一种完全无模型策略梯度强化学习算法, 实现了马尔科夫随机系统动力学完全未知情况下的最优控制. 最后, 通过仿真结果证明了本文所提两种无模型策略梯度强化学习最优控制算法的高效性与优越性.

关键词 随机系统, 最优控制, 未知动力学, 策略梯度, 强化学习

引用格式 杜城龙, 韩洁, 李繁飙, 桂卫华. 基于无模型策略梯度强化学习的未知随机系统最优控制. 自动化学报, 2025, **51**(10): 2245−2255

DOI 10.16383/j.aas.c250156 **CSTR** 32138.14.j.aas.c250156

Model-free Policy Gradient-based Reinforcement Learning Algorithms for Optimal Control of Unknown Stochastic Systems

DU Cheng-Long[1] HAN Jie[1] LI Fan-Biao[1] GUI Wei-Hua[1]

Abstract This paper investigates the optimal control problem of a class of Markov stochastic jump systems (MSJSs) with unknown dynamics by two novel model-free policy gradient (PG)-based reinforcement learning (RL) algorithms. Firstly, for MSJSs with partially unknown model information, an analytical form of model-free PG is derived based on the sampling data of MSJSs and the solutions to coupled Lyapunov equations, and a partially model-free PG-based RL optimal control algorithm is proposed, where the predefined performance index is directly minimized. As the fact that the necessary data for solving the coupled Lyapunov equations and calculating the PG can be extracted from the same trajectory of the system sampling data, without the need to collect additional sampling data, the sampling complexity of the algorithm is significantly reduced. Furthermore, in order to completely eliminate the dependence on the model information of MSJSs, the PG is estimated through random perturbation feedback gain, and a completely model-free PG-based RL algorithm is proposed to achieve optimal control of MSJSs with completely unknown dynamics. Finally, simulation results are presented to demonstrate the efficiency and superiority of the proposed two model-free PG-based RL optimal control algorithms.

Key words Stochastic systems, optimal control, unknown dynamics, policy gradient, reinforcement learning

Citation Du Cheng-Long, Han Jie, Li Fan-Biao, Gui Wei-Hua. Model-free policy gradient-based reinforcement learning algorithms for optimal control of unknown stochastic systems. *Acta Automatica Sinica*, 2025, **51**(10): 2245−2255

实际工业应用中系统不确定性普遍存在, 包括环境扰动、参数摄动、结构变化、工况迁移等, 这些不确定性往往具有随机特性[1], 如车辆系统[2]、电力系统[3] 和复杂网络[4] 等, 由此随机系统得到了广泛研究与应用. 与确定性系统相比, 随机系统由于其内含的随机因素, 表现出更为复杂的动态行为[5], 对系统建模、稳定性分析与控制设计提出更高的要求.

马尔科夫随机系统作为一类典型的随机系统, 被广泛应用于描述系统参数衍变、工况迁移、操作模式变化等结构随机突变的复杂特性, 即系统依据马尔科夫链在多个运行模态之间相互切换的过程, 在汽车工业、航空航天、智慧交通、金融经济等多个领域得到深入探索和应用[6−7]. 与此同时, 马尔科夫随机系统的稳定性分析、镇定设计和鲁棒控制作为

收稿日期 2025-04-15 录用日期 2025-06-19
Manuscript received April 15, 2025; accepted June 19, 2025
国家自然科学基金 (62303492, 62533005, 62222317, 62473383), 湖南省自然科学基金 (2025JJ40056, 2023JJ40765), 广东省基础与应用基础研究基金 (2024A1515240069) 资助
Supported by National Natural Science Foundation of China (62303492, 62533005, 62222317, 62473383), Natural Science Foundation of Hunan Province (2025JJ40056, 2023JJ40765), and Guangdong Basic and Applied Basic Research Foundation (2024A1515240069)
本文责任编委 温广辉
Recommended by Associate Editor WEN Guang-Hui
1. 中南大学自动化学院 长沙 410083
1. School of Automation, Central South University, Changsha 410083

该领域的基本问题也得到深入的研究[8−11]. 近年来, 一批新的研究成果不断涌现: 如 Zhang 等[12] 提出一种基于积分强化学习的马尔科夫随机系统鲁棒控制算法; Tian 等[13] 提出针对部分已知转移概率和非线性因素的马尔科夫随机系统的鲁棒控制方法; 文献 [14−15] 设计一种新型的滑模函数, 并通过合成离散时间滑模控制律保证了马尔科夫随机系统的有限时间稳定; 文献 [16−18] 给出马尔科夫随机系统的随机稳定性判定准则; 随后, 文献 [19−20] 考虑具有双重切换特性的马尔科夫随机系统, 给出有限时间稳定性判定准则和鲁棒积分滑模镇定设计方法, 进一步推广至离散马尔科夫随机系统并给出指数稳定性判定准则及输出反馈异步滑模控制方法.

在此基础上, 最优控制问题作为马尔科夫随机系统控制研究的重要分支, 旨在实现系统稳定运行的同时优化其控制性能[21−22]. 为最小化马尔科夫随机系统的预定义最优性能函数, 研究人员提出了多种优化方法, 如无求逆的迭代算法 (Inversion-free iterative algorithm)[21]、随机动态规划方法 (Stochastic dynamic programming method)[22] 以及基于策略梯度的强化学习 (Policy gradient-based reinforcement learning) 算法[23−24]. 在这些方法中, 基于策略梯度 (Policy gradient, PG) 的强化学习算法凭借其算法简洁和适用性广的优势, 可实现直接在参数空间搜索最优控制增益, 已被广泛应用于马尔科夫随机系统的最优控制协议设计. 文献 [23] 研究了策略梯度方法在马尔科夫随机系统的二次最优控制问题中的全局收敛性. 文献 [24] 进一步分析了 Gauss-Newton 策略梯度方法和 Natural 策略梯度方法在马尔科夫随机系统最优控制中的线性收敛性. 需要强调的是, 这些研究结果均依赖于马尔科夫随机系统的准确模型信息, 属于基于模型的策略梯度强化学习算法[23−24]. 然而, 由于复杂的运行环境、多模态跳变、未建模动力学及其他因素[25−26], 在实际场景中获取马尔科夫随机系统的准确模型信息存在巨大挑战.

近年来, 基于学习的控制方法, 如策略迭代 (Policy iteration, PI)、值迭代 (Value iteration, VI)、策略梯度, 已被广泛应用于线性系统[27−28] 和非线性系统[29−32] 的研究中. 这些基础研究极大地推动了该领域的发展. 需要注意的是, 这些系统均为确定性系统, 不涉及随机因素. 相比之下, 复杂的随机动力学特性[5] 使得现有针对确定性系统的基于学习的控制方法不再适用. 为实现马尔科夫随机系统的最优控制, 同时避免对模型信息的依赖, 学者们提出了无导数的策略梯度方法, 尤其是零阶优化算法[27, 33]. 这些方法通过对反馈增益进行随机摄动来估计策略梯度, 从而避免对模型信息的依赖. 需要注意的是, 因为估计策略梯度需要完整的系统轨迹, 所以这种基于随机摄动的方法会增加一定的采样复杂度. 为此, 本文提出一种部分无模型策略梯度强化学习最优控制算法, 通过马尔科夫随机系统输入输出轨迹数据和耦合 Lyapunov 方程解, 得到策略梯度的解析形式, 降低了采样复杂度. 其中, 耦合 Lyapunov 方程的解无需额外采集数据即可获得, 由此在显著降低数据采样率的同时通过学习获得了马尔科夫随机系统的最优控制协议. 此外, 在模型信息完全未知的情况下, 通过对反馈增益进行随机摄动估计策略梯度, 提出完全无模型的策略梯度强化学习最优控制算法, 彻底避免了对所有模型信息的依赖.

本文的主要贡献总结如下:

1) 针对部分未知动力学马尔科夫随机系统的最优控制问题, 提出一种具有高采样效率的部分无模型策略梯度强化学习算法. 基于马尔科夫随机系统采样数据和耦合 Lyapunov 方程解, 推导得到强化学习策略梯度的显式解析形式, 显著降低了采样复杂度.

2) 所提部分无模型策略梯度强化学习算法中的耦合 Lyapunov 方程的解, 可以通过无模型的方式迭代求解解耦后的 Lyapunov 方程获得. 与传统方法相比, 本算法仅需利用马尔科夫随机系统相同轨迹数据即可同时求解 Lyapunov 方程解和策略梯度, 而在执行所提算法时无需采集额外的采样数据.

3) 进一步提出一种完全无模型策略梯度强化学习的最优控制算法, 即, 在模型信息完全未知的情况下, 通过反馈增益随机摄动估计强化学习的策略梯度, 彻底避免了对所有模型信息的依赖. 与部分无模型策略梯度强化学习算法相比, 完全无模型策略梯度强化学习算法提供了一种基于摄动采样的数值方法, 有望应用于更复杂的场景.

1 问题描述

考虑一类马尔科夫随机系统:

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

其中, $x(t) \in \mathbf{R}^n$ 和 $u(t) \in \mathbf{R}^m$ 分别表示系统状态和控制输入; $\{r(t),\ t \geq 0\}$ 是一个取值于有限集合 $\mathcal{N} = \{1,\ 2,\ \cdots,\ N\}$ 的马尔科夫过程, 其生成矩阵为 $\Pi = \{\pi_{ij}\}$, $i,\ j \in \mathcal{N}$; 本文其他符号定义可参考表 1.

从时刻 t 的模态 i 跳变到时刻 $t+\delta$ 的模态 j 的

表 1　符号说明
Table 1　Notations explanation

符号	说明
$\mathbf{R}^m$, $\mathbf{R}^{m\times n}$	实数向量, 实数矩阵
$\mathrm{E}\{\cdot\}$	数学统计期望
$\otimes$	克罗内克 (Kronecker) 积
$\mathbf{Z}$, $\mathbf{Z}_{\geq 0}$	整数集, 非负整数集
$X>\mathbf{0}$, $X\geq\mathbf{0}$	正定矩阵 X, 半正定矩阵 X
I_n, $\mathbf{0}$	单位矩阵, 零矩阵
$\mathrm{vec}(X)$	$[x_1^{\mathrm{T}},\ \cdots,\ x_m^{\mathrm{T}}]^{\mathrm{T}}\in\mathbf{R}^{n\times m}$, $X\in\mathbf{R}^{n\times m}$
$\mathrm{vecs}(X)$	$[x_{11},\ \cdots,\ x_{1n},\ x_{22},\ \cdots,\ x_{2n},\ \cdots,\ x_{nn}]^{\mathrm{T}}\in\mathbf{R}^{\frac{n(n+1)}{2}}$, $X\in\mathbf{R}^{n\times n}$
$\bar{x}$	$[x_1^2,\ \cdots,\ 2x_1x_n,\ x_2^2,\ \cdots,\ 2x_2x_n,\ \cdots,\ x_n^2]^{\mathrm{T}}\in\mathbf{R}^{\frac{n(n+1)}{2}}$, $x=[x_1,\ \cdots,\ x_n]^{\mathrm{T}}\in\mathbf{R}^n$

转移概率为:

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

其中, $\delta>0$, $\lim_{\delta\to 0}o(\delta)/\delta=0$, 且对于 $j\neq i$, 有 $\pi_{ij}>0$, 而 $\pi_{ii}=-\sum_{j=1,\,j\neq i}^{N}\pi_{ij}$, $i,\ j\in\mathcal{N}$. 矩阵 $A_{r(t)}\in\mathbf{R}^{n\times n}$ 和 $B_{r(t)}\in\mathbf{R}^{n\times m}$ 分别表示系统模态 $r(t)$ 的系统矩阵和输入矩阵. 为简化表示, 取 $r(t)=i$, 则可将 (1) 重写为:

$$\dot{x}(t)=A_ix(t)+B_iu(t)\tag{3}$$

引入马尔科夫随机系统 (1) 的最优控制定义如下.

定义 1[34]**.** 考虑具有转移概率 (2) 的马尔科夫随机系统 (1), 对于任意初始有限状态 $x(0)$, 若性能指标函数

$$\begin{aligned}&J(x(0)\in\mathcal{D}_0,\ r(0)\in\rho_0)=\\&\quad\mathrm{E}\Bigg\{\int_0^{\infty}\big(x^{\mathrm{T}}(\tau)Q_ix(\tau)+\\&\quad u^{\mathrm{T}}(\tau)R_iu(\tau)\big)\mathrm{d}\tau\Big|x(0)\in\mathcal{D}_0,\ r(0)\in\rho_0\Bigg\}\end{aligned}\tag{4}$$

被最小化至 $J^*(x(0)\in\mathcal{D}_0,\ r(0)\in\rho_0)$, 其中 $\mathcal{D}_0$ 和 ρ_0 分别是初始状态和模态的分布, 且 $Q_i\geq\mathbf{0}$, $R_i>\mathbf{0}$, 则实现了马尔科夫随机系统 (1) 的最优控制.

接下来, 给出一些假设和引理.

假设 1. 马尔科夫过程 $r(t)$ 是可遍历的.

假设 2. 具有转移概率 (2) 的马尔科夫随机系统 (1) 可镇定.

假设 3. 转移概率是已知的.

引理 1[34]**.** 对于马尔科夫随机系统 (1), 给定 $Q_i\geq\mathbf{0}$, $R_i>\mathbf{0}$, 最优控制协议为:

$$u(t)=K_i^*x(t),\quad K_i^*=-R_i^{-1}B_i^{\mathrm{T}}P_i^*\tag{5}$$

该协议可以解决定义 1 中的问题, 即可以得到最优的 $J^*(x(0)\in\mathcal{D}_0,\ r(0)\in\rho_0)$. 其中, $P_i^*>\mathbf{0}$ 是非线性耦合代数 Riccati 方程 (Coupled algebraic Riccati equations, CAREs) 的唯一解:

$$\begin{aligned}&P_i^*A_i+A_i^{\mathrm{T}}P_i^*-P_i^*B_iR_i^{-1}B_i^{\mathrm{T}}P_i^*+\\&Q_i+\sum_{j=1}^{N}\pi_{ij}P_j^*=\mathbf{0}\end{aligned}\tag{6}$$

引理 2[35]**.** 考虑马尔科夫随机系统 (1), 如果给定的 $Q_i\geq\mathbf{0}$, $R_i>\mathbf{0}$ 和 K_i, $i\in\mathcal{N}$, 使得马尔科夫随机系统 (1) 是均方稳定的, 即, 对于任意初始有限状态 $x(0)$ 和初始模态 $r(0)$, 满足 $\lim_{t\to\infty}\mathrm{E}\{x(t)x^{\mathrm{T}}(t)|\ x(0)\in\mathcal{D}_0,\ r(0)\in\rho_0\}=0$, 则存在唯一的正定矩阵 P_i, $i\in\mathcal{N}$, 使得以下耦合 Lyapunov 方程成立:

$$\begin{aligned}&P_i(\bar{A}_i+B_iK_i)+(\bar{A}_i+B_iK_i)^{\mathrm{T}}P_i+K_i^{\mathrm{T}}R_i\,K_i+\\&Q_i+\sum_{j=1,\,j\neq i}^{N}\pi_{ij}P_j=\mathbf{0}\end{aligned}\tag{7}$$

其中, $\bar{A}_i=A_i+\frac{\pi_{ii}}{2}I_n$.

由于复杂的运行环境和多模态跳变等问题, 很难对马尔科夫随机系统 (1) 的动力学进行准确建模[25−26]. 本文旨在提出基于策略梯度的无模型强化学习算法, 以解决未知动力学马尔科夫随机系统 (1) 的最优控制问题.

2　主要结果

本部分主要提出两种基于策略梯度的无模型强化学习算法, 获取部分或完全未知动力学马尔科夫随机系统的最优控制策略. 首先, 针对系统矩阵 A_i 未知、输入矩阵 B_i 已知的马尔科夫随机系统, 提出一种部分无模型策略梯度强化学习的最优控制算法. 该算法提供了策略梯度的解析形式, 极大地提高了采样效率. 随后, 针对系统矩阵 A_i、输入矩阵 B_i 均未知的马尔科夫随机系统, 提出一种完全无模型策略梯度强化学习的最优控制算法. 该算法通过随机摄动反馈增益估计策略梯度, 可以完全解除对模型信息的依赖.

2.1　部分无模型策略梯度强化学习最优控制算法

本节提出一种部分无模型策略梯度强化学习最优控制算法, 解决系统矩阵未知、输入矩阵已知的马尔科夫随机系统 (1) 的最优控制问题. 所推导的策略梯度由马尔科夫随机系统 (1) 采集的样本数据和耦合 Lyapunov 方程解的解析函数所表示, 从而提高算法的采样效率.

对于未知系统矩阵的马尔科夫随机系统 (1), 其控制器的形式设计为

$$u(t) = K_i(k)x(t) \tag{8}$$

其中, $i \in \mathcal{N}$, k 是迭代索引. 在提出部分无模型策略梯度强化学习算法之前, 为第 i 个模态定义以下矩阵:

$$D_{i1} = \left[\bar{x}(\tau)|_{t^i_{0-}}^{t^i_{0+}}, \cdots, \bar{x}(\tau)|_{t^i_{l-}}^{t^i_{l+}}\right]^{\mathrm{T}}$$

$$D_{i2} = \left[\int_{t^i_{0-}}^{t^i_{0+}} \bar{x}(\tau)\mathrm{d}\tau, \cdots, \int_{t^i_{l-}}^{t^i_{l+}} \bar{x}(\tau)\mathrm{d}\tau\right]^{\mathrm{T}}$$

$$D_{i3} = \left[\int_{t^i_{0-}}^{t^i_{0+}} (x(\tau)\otimes x(\tau))\mathrm{d}\tau, \cdots, \int_{t^i_{l-}}^{t^i_{l+}} (x(\tau)\otimes x(\tau))\mathrm{d}\tau\right]^{\mathrm{T}}$$

$$D_{i4} = D_{i1} + \pi_{ii}D_{i2}$$

$$\Delta_{i\iota} = \int_{t^i_{\iota-}}^{t^i_{\iota-}+T_s} (R_i u(t^i_{\iota-}) + B_i^{\mathrm{T}}P_i(k)x(t^i_{\iota-}))x^{\mathrm{T}}(\tau)\mathrm{d}\tau$$

$$\Delta_i = \left[\Delta_{i0}, \Delta_{i1}, \cdots, \Delta_{il}\right]$$

其中, $l \in \mathbf{Z}$, $0 \le t^i_{0-} < t^i_{0+} \le t^i_{1-} < t^i_{1+} \le \cdots \le t^i_{l-} < t^i_{l+}$; $T_s > 0$ 为采样周期. 基于文献 [28, 36−37], 学习率 γ_k 需要满足

$$\gamma_k > 0, \ \sum_{k=0}^{\infty}\gamma_k = \infty, \ \sum_{k=0}^{\infty}(\gamma_k)^2 < \infty \tag{9}$$

随后, 设计部分无模型策略梯度强化学习算法, 如算法 1 所示, 该算法用于学习未知系统矩阵马尔科夫随机系统 (1) 的最优控制增益 K_i^*, $i \in \mathcal{N}$.

注 1. 传统基于策略梯度的强化学习算法[27, 33] 通常利用随机摄动反馈增益获取其策略梯度, 而算法 1 则通过直接利用马尔科夫随机系统采样数据和耦合 Lyapunov 方程解获取策略梯度的解析形式, 即 $\int_{t^i_{\iota-}}^{t^i_{\iota-}+T_s}(R_i u(t^i_{\iota-}) + B_i^{\mathrm{T}}P_i(k)x(t^i_{\iota-}))x^{\mathrm{T}}(\tau)\mathrm{d}\tau$. 因此, 该算法能够在降低采样复杂度的同时得到最优控制增益.

注 2. 从算法 1 的第 4 行和第 13 行可以看出, 求解解耦后的 Lyapunov 方程所需的采样数据, 即 D_{i1}, D_{i2}, D_{i3}, D_{i4}, 以及计算策略梯度所需的信息, 即 Δ_i, 均可以从马尔科夫随机系统 (1) 的相同轨迹数据中提取, 而无需采集额外的数据. 此外, 数据在第 6 行到第 11 行的循环之前收集, 所采集的数据可以在第 6 行到第 11 行的循环中重复使用. 因此, 算法 1 具有更高的数据利用率.

算法 1. 部分无模型策略梯度强化学习算法

1: 初始化: 选择初始均方稳定的控制增益 $K_i(0)$, $i \in \mathcal{N}$, $\gamma_k > 0$, $N_s > 0$, $T_t > 0$, $T_s > 0$ 以及 $\epsilon > 0$;
2: **for** $k = 0 \to \infty$ **do**
3: 通过对马尔科夫随机系统 (1) 施加 $u(t) = K_i(k)x(t)$, 获取运行时长为 T_t 的 N_s 条采样轨迹;
4: 从采集到的 N_s 条轨迹中提取 D_{i1}, D_{i2}, D_{i3}, D_{i4};
5: $P_i(k, 0) = \mathbf{0}$;
6: **for** $h = 1 \to \infty$ **do**
7: 计算 $\mathrm{vecs}(P_i(k, h)) = -(D_{i4}^{\mathrm{T}}D_{i4})^{-1}D_{i4}^{\mathrm{T}}D_{i3} \times \mathrm{vec}(Q_i(k, h))$, 得到 $P_i(k, h)$, 其中 $Q_i(k, h) = K_i^{\mathrm{T}}(k)R_iK_i(k) + Q_i + \sum_{j=1}^{i-1}\pi_{ij}P_j(k, h) + \sum_{j=i+1}^{N}\pi_{ij}P_j(k, h-1)$;
8: **if** $h \ge 1$ 且 $\|P_i(k, h) - P_i(k, h-1)\| \le \epsilon$, $\forall i \in \mathcal{N}$ **then**
9: $h_T = h$, 并跳出循环;
10: **end if**
11: **end for**
12: $P_i(k) = P_i(k, h_T)$;
13: 从收集到的 N_s 条轨迹数据中提取 Δ_i;
14: **for** $\iota = 0 \to l$ **do**
15: 根据 $K_i(k+1) = K_i(k) - 2\frac{\gamma_k}{l}\Delta_{i\iota}$ 更新 $K_i(k+1)$;
16: **if** $\|K_i(k+1) - K_i(k)\| \le \epsilon$, $\forall i \in \mathcal{N}$ **then**
17: 跳出循环;
18: **end if**
19: **end for**
20: **end for**
21: 返回 $K_i(k+1)$.

注 3. 关于数据需求量、学习精度和计算负担之间的关系, 从算法 1 可以明显看出, 当 T_s 减小时, 采集的数据量增加, 可提高学习精度. 然而, 数据量的增加也将显著增加其计算负担. 相反, 当 T_s 增大时, 采集的数据量减少, 将导致学习精度降低, 但计算负担将大幅减轻. 因此, 通过经验选择一个合适的采样间隔 T_s 至关重要, 该间隔不仅要满足学习精度的要求, 还要确保计算和采样资源的合理分配.

定理 1. 考虑满足假设 1 ~ 3 的马尔科夫随机系统 (1), 控制器设计为 (8), 给定初始均方稳定的控制增益 $K_i(0)$, $i \in \mathcal{N}$, 学习率 γ_k 满足条件 (9), 那么, 可以通过部分无模型策略梯度强化学习算法 1 学习到最优控制协议 (5), 且性能函数严格单调递减, 对于所有 $i \in \mathcal{N}$, 满足 $\lim_{k\to\infty}K_i(k) = K_i^*$, $\lim_{k\to\infty}P_i(k) = P_i^*$, $\lim_{k\to\infty}J_k(x(t) \in \mathcal{D}_t, r(t) \in \rho_t) = J^*(x(t) \in \mathcal{D}_t, r(t) \in \rho_t)$.

证明. 首先, 关注 $J_k(x(t) \in \mathcal{D}_t, r(t) \in \rho_t)$ 的单

调递减特性.

步骤 1. 选择控制增益 $K_i(0)$, $i\in\mathcal{N}$, 使得 (1) 是均方稳定的.

步骤 k. 当 $k>0$ 时, 假设迭代控制增益 $K_i(k)$, $i\in\mathcal{N}$ 能够确保马尔科夫随机系统 (1) 的均方稳定性, 接下来将证明在第 $k+1$ 次迭代时, 迭代控制增益 $K_i(k+1)$, $i\in\mathcal{N}$ 仍然能够保证马尔科夫随机系统 (1) 的均方稳定性.

根据引理 2, 在控制增益 $K_i(k)$, $i\in\mathcal{N}$ 均方稳定的条件下, 以下耦合 Lyapunov 方程具有唯一解 $P_i(k)$, $i\in\mathcal{N}$:

$$\begin{aligned}&P_i(k)(\bar{A}_i+B_iK_i(k))+(\bar{A}_i+B_iK_i(k))^{\mathrm{T}}P_i(k)+\\&K_i^{\mathrm{T}}(k)R_iK_i(k)+Q_i+\sum_{j=1,\,j\neq i}^{N}\pi_{ij}P_j(k)=\mathbf{0}\end{aligned}\tag{10}$$

此外, 对于马尔科夫随机系统 (1), 有如下关系成立:

$$\begin{aligned}&\mathrm{E}\{x^{\mathrm{T}}(t)P_i(k)x(t)|x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t\}=\\&\quad\mathrm{E}\Big\{\int_t^{\infty}x^{\mathrm{T}}(\tau)(Q_i+K_i^{\mathrm{T}}(k)R_iK_i(k))x(\tau)\mathrm{d}\tau\Big|\\&\quad x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t\Big\}\end{aligned}\tag{11}$$

其中, $\mathcal{D}_t$ 和 ρ_t 分别是时刻 t 处状态 $x(t)$ 和模态 $r(t)$ 的分布. 令

$$\begin{aligned}&V_k(x(t)|x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t)=\\&\quad\mathrm{E}\Big\{\int_t^{\infty}x^{\mathrm{T}}(\tau)(Q_i+K_i^{\mathrm{T}}(k)R_iK_i(k))x(\tau)\mathrm{d}\tau\Big|\\&\quad x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t\Big\}\end{aligned}\tag{12}$$

$$\begin{aligned}&Q_k(x(t)|x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t)=\\&\quad\mathrm{E}\Big\{\int_t^{t+T_s}(x^{\mathrm{T}}(\tau)Q_ix(\tau)+u^{\mathrm{T}}(\tau)R_iu(\tau))\mathrm{d}\tau\Big|\\&\quad x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t\Big\}+V_k(x(t+T_s)|x(t+\\&\quad T_s)\in\mathcal{D}_{t+T_s},\ r(t+T_s)\in\rho_{t+T_s})\end{aligned}\tag{13}$$

其中, $\mathcal{D}_{t+T_s}$ 和 ρ_{t+T_s} 分别是时刻 $t+T_s$ 处状态 $x(t+T_s)$ 和模态 $r(t+T_s)$ 的分布. 由 (11) 和 (12) 可得

$$\begin{aligned}&V_k(x(t)+\Phi|x(t+T_s)\in\mathcal{D}_{t+T_s},\ r(t+T_s)\in\rho_{t+T_s})=\\&\quad\mathrm{E}\{x^{\mathrm{T}}(t)P_i(k)x(t)|x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t\}+\\&\quad\mathrm{E}\{2x^{\mathrm{T}}(t)P_i(k)\Phi|x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t\}+\\&\quad\mathrm{E}\{\Phi^{\mathrm{T}}P_i(k)\Phi|x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t\}\end{aligned}\tag{14}$$

定义变量 $\Phi:=\int_t^{t+T_s}(A_ix(\tau)+B_iu(\tau))\mathrm{d}\tau$, 可将 $Q_k(x(t)|x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t)$ 重写为

$$\begin{aligned}&Q_k(x(t)|x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t)=\\&\quad\mathrm{E}\Big\{\int_t^{t+T_s}(x^{\mathrm{T}}(\tau)Q_ix(\tau)+u^{\mathrm{T}}(\tau)R_iu(\tau))\mathrm{d}\tau\Big|\\&\quad x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t\Big\}+\\&\quad\mathrm{E}\{x^{\mathrm{T}}(t)P_i(k)x(t)|x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t\}+\\&\quad\mathrm{E}\{2x^{\mathrm{T}}(t)P_i(k)\Phi|x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t\}+\\&\quad\mathrm{E}\{\Phi^{\mathrm{T}}P_i(k)\Phi|x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t\}\end{aligned}\tag{15}$$

当 T_s 足够小时, 有 $\frac{\Phi^{\mathrm{T}}P_i(k)\Phi}{2x^{\mathrm{T}}(t)P_i(k)\Phi}\approx\mathbf{0}$. 因此, 可以得到

$$\begin{aligned}&Q_k(x(t)|x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t)\approx\\&\quad\mathrm{E}\Big\{\int_t^{t+T_s}(x^{\mathrm{T}}(\tau)Q_ix(\tau)+u^{\mathrm{T}}(\tau)R_iu(\tau))\mathrm{d}\tau\Big|\\&\quad x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t\Big\}+\\&\quad\mathrm{E}\{x^{\mathrm{T}}(t)P_i(k)x(t)|x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t\}+\\&\quad\mathrm{E}\{2x^{\mathrm{T}}(t)P_i(k)\Phi|x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t\}\end{aligned}\tag{16}$$

那么 $J_k(x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t)$ 可以沿着以下策略梯度进行优化:

$$\begin{aligned}&\frac{\partial J_k(x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t)}{\partial K_i(k)}=\\&\quad\frac{\partial Q_k(x(t)|x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t)}{\partial u(\tau)}\frac{\partial u(\tau)}{\partial K_i(k)}\end{aligned}\tag{17}$$

其中

$$\begin{aligned}&J_k(x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t)=\\&\quad\mathrm{E}\Big\{\int_t^{\infty}\big(x^{\mathrm{T}}(\tau)Q_ix(\tau)+u^{\mathrm{T}}(\tau)R_iu(\tau)\big)\mathrm{d}\tau\Big|\\&\quad x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t\Big\}\end{aligned}\tag{18}$$

根据 (16), 选取足够小的 T_s, 可以得到

$$\begin{aligned}&\frac{\partial Q_k(x(t)|x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t)}{\partial u(\tau)}\frac{\partial u(\tau)}{\partial K_i(k)}\approx\\&\quad\mathrm{E}\Big\{\int_t^{t+T_s}\Big(\frac{\partial(u^{\mathrm{T}}(\tau)R_iu(\tau))}{\partial u(\tau)}+\\&\quad2\frac{\partial(x^{\mathrm{T}}(t)P_i(k)B_iu(\tau))}{\partial u(\tau)}\Big)\times\end{aligned}$$

$$\frac{\partial u(\tau)}{\partial K_i(k)}\mathrm{d}\tau\Big|x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t\Big\}\approx$$
$$2\mathrm{E}\Big\{\int_t^{t+T_s}(R_iu(t)+B_i^{\mathrm{T}}P_i(k)x(t))x^{\mathrm{T}}(\tau)\mathrm{d}\tau\Big|$$
$$x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t\Big\} \tag{19}$$

其中, $P_i(k)$ 通过求解 (10) 得到. 更具体地说, 基于文献 [38], 耦合 Lyapunov 方程 (10) 的解可以通过迭代求解以下解耦 Lyapunov 方程获得:

$$P_i(k,\ h)(\bar{A}_i+B_iK_i(k))+(\bar{A}_i+B_iK_i(k))^{\mathrm{T}}P_i(k,\ h)+$$
$$K_i^{\mathrm{T}}(k)R_iK_i(k)+Q_i+\sum_{j=1}^{i-1}\pi_{ij}P_j(k,\ h-1)+$$
$$\sum_{j=i+1}^{N}\pi_{ij}P_j(k,\ h-1)=\mathbf{0} \tag{20}$$

其中, $P_i(k,\ 0)=\mathbf{0}$; h 为迭代索引. 为实现比 (20) 更快的收敛速度, 根据文献 [35] 中的方法提出一种修改方案, 将上一次迭代的 $\pi_{ij}P_j(k,\ h-1)$ 替换为更新后的 $\pi_{ij}P_j(k,\ h)$, 其中 $j<i$. 将 (20) 修改为如下形式:

$$P_i(k,\ h)(\bar{A}_i+B_iK_i(k))+(\bar{A}_i+B_iK_i(k))^{\mathrm{T}}P_i(k,\ h)+$$
$$K_i^{\mathrm{T}}(k)R_iK_i(k)+Q_i+\sum_{j=1}^{i-1}\pi_{ij}P_j(k,\ h)+$$
$$\sum_{j=i+1}^{N}\pi_{ij}P_j(k,\ h-1)=\mathbf{0} \tag{21}$$

其中, $P_i(k,\ 0)=\mathbf{0}$. 需要注意的是, 使用最新获得的信息 $\sum_{j=1}^{i-1}\pi_{ij}P_j(k,\ h)$ 代替上一次迭代的 $\sum_{j=1}^{i-1}\pi_{ij}\times P_j(k,\ h-1)$, 因此实现了更快的收敛速度. 根据 (21) 可得:

$$\frac{\mathrm{d}}{\mathrm{d}t}(x^{\mathrm{T}}(t)P_i(k,\ h)x(t))=$$
$$-x^{\mathrm{T}}(t)(\pi_{ii}P_i(k,\ h)+K_i^{\mathrm{T}}(k)R_iK_i(k)+Q_i+$$
$$\sum_{j=1}^{i-1}\pi_{ij}P_j(k,\ h)+\sum_{j=i+1}^{N}\pi_{ij}P_j(k,\ h-1))x(t) \tag{22}$$

对上述方程的两侧进行积分, 可得:

$$(x^{\mathrm{T}}(\tau)\otimes x^{\mathrm{T}}(\tau))|_t^{t+\kappa}\,\mathrm{vec}(P_i(k,\ h))=$$
$$-\pi_{ii}\int_t^{t+\kappa}(x^{\mathrm{T}}(\tau)\otimes x^{\mathrm{T}}(\tau))\mathrm{d}\tau\,\mathrm{vec}(P_i(k,\ h))-$$
$$\int_t^{t+\kappa}(x^{\mathrm{T}}(\tau)\otimes x^{\mathrm{T}}(\tau))\mathrm{d}\tau\,\mathrm{vec}(Q_i(k,\ h)) \tag{23}$$

其中,

$$Q_i(k,\ h)=K_i^{\mathrm{T}}(k)R_iK_i(k)+Q_i+\sum_{j=1}^{i-1}\pi_{ij}P_j(k,\ h)+$$
$$\sum_{j=i+1}^{N}\pi_{ij}P_j(k,\ h-1) \tag{24}$$

$\kappa>0$ 为采样间隔. 那么, (23) 可以被重写为更紧凑的形式:

$$D_{i4}\mathrm{vecs}(P_i(k,\ h))=-D_{i3}\mathrm{vec}(Q_i(k,\ h)) \tag{25}$$

通过对具有学习控制器形式 $u(t)=K_i(k)x(t)$ 的马尔科夫随机系统 (1) 进行充分的数据采集, 可以保证 $(D_{i4}^{\mathrm{T}}D_{i4})^{-1}$ 的存在[28]. 于是, 如下关系成立:

$$\mathrm{vecs}(P_i(k,\ h))=-(D_{i4}^{\mathrm{T}}D_{i4})^{-1}D_{i4}^{\mathrm{T}}D_{i3}\mathrm{vec}(Q_i(k,\ h)) \tag{26}$$

因此, 耦合 Lyapunov 方程 (10) 的解可以通过迭代求解解耦 Lyapunov 方程 (21) 获得, 而这些方程是通过 (25) 以无模型的方式求解的. 最终, 控制增益 $K_i(k+1)$, $i\in\mathcal{N}$ 更新为:

$$K_i(k+1)=$$
$$K_i(k)-\gamma_k\frac{\partial Q_k(x(t)|x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t)}{\partial K_i(k)}\approx$$
$$K_i(k)-2\gamma_k\mathrm{E}\Big\{\int_t^{t+T_s}(R_iu(t)+$$
$$B_i^{\mathrm{T}}P_i(k)x(t))x^{\mathrm{T}}(\tau)\mathrm{d}\tau\Big|x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t\Big\} \tag{27}$$

然后, 可以得到:

$$J_{k+1}(x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t)\le$$
$$J_k(x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t)<\infty \tag{28}$$

其中, $J_k(x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t)$ 是有限的, 即 $J_k(x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t)<\infty$. 因此在第 k 次迭代时, 迭代控制增益 $K_i(k)$, $i\in\mathcal{N}$ 可以保证马尔科夫随机系统 (1) 的均方稳定性. 根据 (28) 可以得出, 在第 $k+1$ 次迭代时, 迭代控制增益 $K_i(k+1)$, $i\in\mathcal{N}$ 仍然可以保证马尔科夫随机系统 (1) 的均方稳定性.

因此, 基于上述分析, 可以得出在第 k 次迭代时, 迭代控制增益 $K_i(k)$, $i\in\mathcal{N}$ 可以保证马尔科夫随机系统 (1) 的均方稳定性. 此外, 在具有初始均方稳定控制增益 $K_i(0)$, $i\in\mathcal{N}$ 和策略梯度 (27) 的情况下, 可以得到对于任意 $k\in\mathbf{Z}_{\ge0}$, 均有 $J_{k+1}(x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t)\le J_k(x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t)<\infty$.

接下来, 证明当 $k\to\infty$ 时, $J_k(x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t)$ 将收敛到最优值 $J^*(x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t)$.

根据 (27), 可以得出:

$$\frac{\partial Q_k(x(t)|x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t)}{\partial u(\tau)}\frac{\partial u(\tau)}{\partial K_i(k)}=\mathbf{0} \tag{29}$$

和下式相等

$$R_iu(t)+B_i^{\mathrm{T}}P_i(k)x(t)=\mathbf{0} \tag{30}$$

即

$$K_i(k)=-R_i^{-1}B_i^{\mathrm{T}}P_i(k) \tag{31}$$

其中, $P_i(k)$ 是 (10) 的唯一解. 将 $K_i(k)=-R_i^{-1}B_i^{\mathrm{T}}\times P_i(k)$ 代入 (10), 可以将 (10) 重写为:

$$P_i(k)A_i+A_i^{\mathrm{T}}P_i(k)-\Xi_i(k)+Q_i+\sum_{j=1}^{N}\pi_{ij}P_j(k)=\mathbf{0} \tag{32}$$

其中, $\Xi_i(k)=P_i(k)B_iR_i^{-1}B_i^{\mathrm{T}}P_i(k)$. 基于引理 1, 耦合代数 Riccati 方程 (32) 具有唯一解 P_i^*, 即 $P_i(k)=P_i^*$. 因此, 当梯度为零时, 有 $K_i(k)=K_i^*$, $P_i(k)=P_i^*$, 且 $J_k(x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t)=J^*(x(t)\in\mathcal{D}_t, r(t)\in\rho_t)$.

因此, 有 $\lim_{k\to\infty}K_i(k)=K_i^*$, $\lim_{k\to\infty}P_i(k)=P_i^*$, 且 $\lim_{k\to\infty}J_k(x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t)=J^*(x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t)$. 至此, 解决了马尔科夫随机系统 (1) 的最优控制问题. □

2.2　完全无模型策略梯度强化学习最优控制算法

本节进一步研究一种完全无模型策略梯度强化学习最优控制算法, 以获取系统矩阵和输入矩阵均未知的马尔科夫随机系统 (1) 的最优控制协议. 该算法的策略梯度通过随机摄动反馈增益估计得到, 无需依赖马尔科夫随机系统 (1) 的任何模型信息. 与部分无模型策略梯度强化学习算法相比, 完全无模型策略梯度强化学习算法在解决更复杂的应用场景方面更具潜力.

在提出完全无模型策略梯度强化学习算法之前, 为第 i 个模态定义以下矩阵:

$$\Omega_{i\iota}=\frac{U_{i\iota}}{2\beta}\Bigg(\int_{t_{\iota-}^i}^{t_{\iota-}^i+T_t}(x^{\mathrm{T}}(\tau)Q_ix(\tau)+u_{i\iota+}^{\mathrm{T}}(\tau)R_iu_{i\iota+}(\tau))\mathrm{d}\tau-\int_{t_{\iota-}^i}^{t_{\iota-}^i+T_t}(x^{\mathrm{T}}(\tau)Q_ix(\tau)+u_{i\iota-}^{\mathrm{T}}(\tau)R_iu_{i\iota-}(\tau))\mathrm{d}\tau\Bigg)$$

$$\Omega_i=\left[\Omega_{i0},\ \Omega_{i1},\ \cdots,\ \Omega_{il}\right]$$

其中, $l\in\mathbf{Z}$; $0\le t_{0-}^i<t_{1-}^i<\cdots<t_{l-}^i$; T_t 是系统的运行时间; $u_{i\iota+}(\tau)=(K_i(k)+\beta U_{i\iota})x(\tau)$; $u_{i\iota-}(\tau)=(K_i(k)-\beta U_{i\iota})x(\tau)$; $U_{i\iota}$ 是一个在 $\mathbf{R}^{m\times n}$ 上均匀分布的随机矩阵且满足 $\|U_{i\iota}\|=1$; β 是一个平滑因子. 类似地, 根据文献 [36], 学习率 γ_k 也需要满足条件 (9).

随后, 设计完全无模型策略梯度强化学习算法, 如算法 2 所示, 用于学习完全未知系统动力学的马尔科夫系统最优控制增益矩阵 K_i^*, $i\in\mathcal{N}$.

注 4. 不同于算法 1 中提出的部分无模型策略梯度强化学习算法, 算法 2 中的策略梯度通过随机摄动反馈增益获取. 算法 2 提供了一种基于数值的方法, 可解决完全未知系统动力学马尔科夫系统的最优控制问题, 有望解决更复杂的应用场景.

算法 2. 完全无模型策略梯度强化学习算法

初始化: 选择初始均方稳定的控制增益 $K_i(0)$, $i\in\mathcal{N}$, $\gamma_k>0$, $N_s>0$, $T_t>0$, $T_s>0$ 以及一个小的 $\epsilon>0$;
for $k=0\to\infty$ **do**
　通过对马尔科夫随机系统 (1) 施加 $u(t)=K_i(k)\times x(t)$, 获取运行时长为 T_t 的 N_s 条采样轨迹;
　通过反馈增益的随机摄动并利用收集的 N_s 条轨迹获取初始状态, 计算 Ω_i, $i\in\mathcal{N}$;
　for $\iota=0\to l$ **do**
　　根据 $K_i(k+1)=K_i(k)-2\frac{\gamma_k}{l}\Omega_{i\iota}$ 更新 $K_i(k+1)$;
　　if $\|K_i(k+1)-K_i(k)\|\le\epsilon$, $\forall i\in\mathcal{N}$ **then**
　　　跳出循环;
　　end if
　end for
end for
返回 $K_i(k+1)$.

定理 2. 考虑满足假设 1 ~ 3 的马尔科夫随机系统 (1), 控制器设计为 (8), 给定初始均方稳定的控制增益 $K_i(0)$, $i\in\mathcal{N}$, 学习率 γ_k 满足条件 (9), 那么, 可以通过完全无模型策略梯度强化学习算法 2 学习到最优控制协议 (5), 且性能函数严格单调递减, 对于所有 $i\in\mathcal{N}$, 满足 $\lim_{k\to\infty}K_i(k)=K_i^*$, $\lim_{k\to\infty}P_i(k)=P_i^*$, $\lim_{k\to\infty}J_k(x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t)=J^*(x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t)$.

证明. 步骤 1. 首先, 选择控制增益 $K_i(0)$, $i\in\mathcal{N}$, 使得 (1) 是均方稳定的.

步骤 k. 当 $k>0$ 时, 假设迭代控制增益 $K_i(k)$, $i\in\mathcal{N}$ 能够确保马尔科夫随机系统 (1) 的均方稳定性, 接下来将证明在第 $k+1$ 次迭代时, 迭代控制增益 $K_i(k+1)$, $i\in\mathcal{N}$ 仍然能够保证马尔科夫随机系统 (1) 的均方稳定性. 令

$$Q_k(x(t)|x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t)=$$
$$\mathrm{E}\Big\{\int_t^{t+\infty}\big(x^{\mathrm{T}}(\tau)Q_ix(\tau)+u^{\mathrm{T}}(\tau)R_iu(\tau)\big)\mathrm{d}\tau\Big|$$
$$x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t\Big\}\tag{33}$$

其中, $\mathcal{D}_t$ 和 ρ_t 分别是时刻 t 处状态 $x(t)$ 和模态 $r(t)$ 的分布. 对于均方稳定的控制增益 $K_i(k)$, $i\in\mathcal{N}$ 以及足够大的 T_t, 显然有:

$$\mathrm{E}\Big\{\int_{t+T_t}^{\infty}(x^{\mathrm{T}}(\tau)Q_ix(\tau)+u^{\mathrm{T}}(\tau)R_iu(\tau))\mathrm{d}\tau\Big|$$
$$x(t+T_t)\in\mathcal{D}_{t+T_t},\ r(t+T_t)\in\rho_{t+T_t}\Big\}\approx 0\tag{34}$$

其中, $\mathcal{D}_{t+T_t}$ 和 ρ_{t+T_t} 分别是时刻 $t+T_t$ 处状态 $x(t+T_t)$ 和模态 $r(t+T_t)$ 的分布. 因此, 可以得到:

$$Q_k(x(t)|x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t)\approx$$
$$\mathrm{E}\Big\{\int_t^{t+T_t}\big(x^{\mathrm{T}}(\tau)Q_ix(\tau)+u^{\mathrm{T}}(\tau)R_iu(\tau)\big)\mathrm{d}\tau\Big|$$
$$x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t\Big\}\tag{35}$$

根据 (35), 可以得到:

$$\frac{\partial Q_k(x(t)|x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t)}{\partial K_i(k)}\approx$$
$$\mathrm{E}\Big\{\frac{U_i}{2\beta}\Big(\int_t^{t+T_t}(x^{\mathrm{T}}(\tau)Q_ix(\tau)+$$
$$u_{i+}^{\mathrm{T}}(\tau)R_iu_{i+}(\tau))\mathrm{d}\tau-$$
$$\int_t^{t+T_t}(x^{\mathrm{T}}(\tau)Q_ix(\tau)+u_{i-}^{\mathrm{T}}(\tau)R_iu_{i-}(\tau))\mathrm{d}\tau\Big)\Big|$$
$$x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t\Big\}\tag{36}$$

其中, $T_t>0$; $u_{i+}(\tau)=(K_i(k)+\beta U_i)x(\tau)$; $u_{i-}(\tau)=(K_i(k)-\beta U_i)x(\tau)$; U_i 是一个在 $\mathbf{R}^{m\times n}$ 上均匀分布的随机矩阵且满足 $\|U_i\|=1$. 因此, 控制增益 $K_i(k+1)$ 通过以下方式更新:

$$K_i(k+1)=$$
$$K_i(k)-\gamma_k\frac{\partial Q_k(x(t)|x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t)}{\partial K_i(k)}\tag{37}$$

根据 (17) 和 (18), 可以推导出 (28). 因此, 由 (37) 可明显看出, 在第 $k+1$ 次迭代时, 迭代控制增益 $K_i(k+1)$, $i\in\mathcal{N}$ 可以确保马尔科夫随机系统 (1) 的均方稳定性.

因此, 类似地, 可以得到: 1) 在第 k 次迭代时, 迭代控制增益 $K_i(k)$, $i\in\mathcal{N}$ 可以确保马尔科夫随机系统 (1) 的均方稳定性; 2) 对于任意 $k\in\mathbf{Z}_{\geq 0}$, 在初始均方稳定的控制增益 $K_i(0)$, $i\in\mathcal{N}$ 和策略梯度 (27) 的情况下, 不等式 $J_{k+1}(x(t)\in\mathcal{D}_t, r(t)\in\rho_t)\leq J_k(x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t)<\infty$ 成立; 3) $\lim_{k\to\infty}K_i(k)=K_i^*$, $\lim_{k\to\infty}P_i(k)=P_i^*$, $\lim_{k\to\infty}J_k(x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t)=J^*(x(t)\in\mathcal{D}_t,\ r(t)\in\rho_t)$. 至此, 解决了马尔科夫随机系统 (1) 的最优控制问题. □

注 5. 算法 1 和算法 2 之间的主要区别在于: 算法 1 是一种采样效率更高的算法; 而算法 2 可实现模型信息完全未知情况下最优控制协议的自主学习. 当系统矩阵 A_i 未知、控制输入矩阵 B_i 已知时, 使用算法 1 可降低采样复杂度, 以更高的采样效率实现马尔科夫随机系统 (1) 的最优控制. 然而, 当系统矩阵和控制输入矩阵的模型信息均未知时, 算法 2 可通过更多的采样数据以完全无模型的方式获得马尔科夫随机系统 (1) 的最优控制协议. 因此, 在算法 1 和算法 2 的实际应用中, 需要在采样效率与模型依赖性之间进行权衡.

3 仿真结果

考虑包含 5 种模态的马尔科夫随机系统 (1), 系统矩阵与控制输入矩阵如下:

$$A_1=\begin{bmatrix}0.15 & 0.10\\ 0 & -0.15\end{bmatrix},\ B_1=\begin{bmatrix}0\\5\end{bmatrix}$$
$$A_2=\begin{bmatrix}-0.15 & 0\\ 0.10 & 0.15\end{bmatrix},\ B_2=\begin{bmatrix}4\\0\end{bmatrix}$$
$$A_3=\begin{bmatrix}0 & 0.12\\ 0 & 0\end{bmatrix},\ B_3=\begin{bmatrix}0\\3\end{bmatrix}$$
$$A_4=\begin{bmatrix}0 & 0\\ 0.13 & 0\end{bmatrix},\ B_4=\begin{bmatrix}1\\4\end{bmatrix}$$
$$A_5=\begin{bmatrix}0 & -0.14\\ 0.11 & 0.25\end{bmatrix},\ B_5=\begin{bmatrix}0\\5\end{bmatrix}$$

以及转移概率生成矩阵

$$\Pi=\begin{bmatrix}-0.32 & 0.32 & 0 & 0 & 0\\ 0 & -0.36 & 0.36 & 0 & 0\\ 0 & 0 & -0.40 & 0.40 & 0\\ 0 & 0 & 0 & -0.44 & 0.44\\ 0.48 & 0 & 0 & 0 & -0.48\end{bmatrix}$$

其中, 假设 1 和假设 2 均得到了满足. 选择 $Q_i=I_2$, $i\in\mathcal{N}$, $R_1=0.8$, $R_2=0.9$, $R_3=1.0$, $R_4=1.1$, $R_5=1.2$. 算法 1 和算法 2 中的详细参数设置为 $\gamma_k=1/(k+1)$, $N_s=10$, $T_t=10\,\mathrm{s}$, $T_s=0.001\,\mathrm{s}$ 以及

$\epsilon_i = 0.001$.

算法 1 在上述配置下的学习结果如图 1 和图 2 所示. 图 1 显示了性能指标的单调递减. 至于图 2 中的 $K_i(k)$, 可以明显看出矩阵准确收敛到 K_i^*. 这一观察结果表明, 对于系统矩阵 A_i 未知、输入矩阵 B_i 已知的马尔科夫随机系统 (1), 其最优控制问题可以通过所提出的部分无模型策略梯度强化学习算法有效解决. 对于算法 2, 在 A_i 和 B_i 均未知的情况下, 仿真结果如图 3 和图 4 所示, 其最终的学习结果与算法 1 接近, 从而说明本文所提出的完全无模型策略梯度强化学习算法, 可有效解决在模型信息完全未知情况下的马尔科夫随机系统最优控制问题, 具有更广泛的应用潜质.

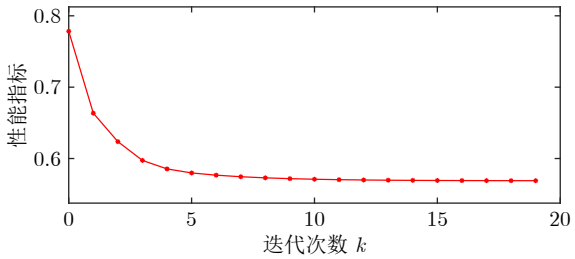


图 1　算法 1 的性能指标

Fig. 1　The performance index of algorithm 1

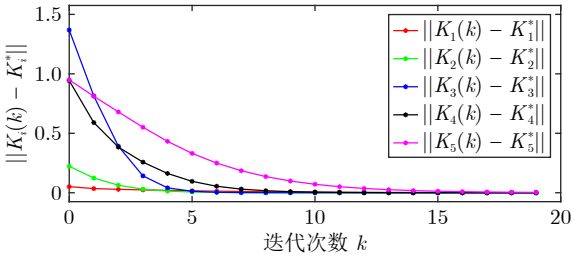


图 2　算法 1 的控制增益矩阵 $K_i(k)$

Fig. 2　The control gain matrix $K_i(k)$ of algorithm 1

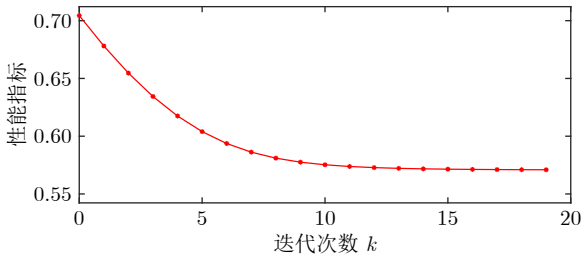


图 3　算法 2 的性能指标

Fig. 3　The performance index of algorithm 2

此外, 为展示算法 1 在提升采样效率方面的优越性, 将其与算法 2 进行比较, 详细信息列在表 2 中. 如表 2 所示, 两种算法均使用前三行中给出的相同参数. 算法 1 的结果与算法 2 的结果高度一致. 然而, 从表 2 的最后一行可以明显看出, 算法 1 的运行时间为 100 s, 而算法 2 的运行时间为 2 000 000 s. 这一比较表明, 算法 1 的采样复杂度得到了显著降低.

通过学习得到的矩阵 K_i^* 和最优控制协议 $u = K_i^* x$ 作用于马尔科夫随机系统 (1), 在图 5 所示的相关模态演化下, 系统状态响应如图 6 所示,

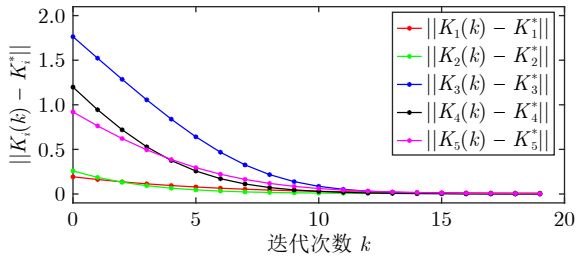


图 4　算法 2 的控制增益矩阵 $K_i(k)$

Fig. 4　The control gain matrix $K_i(k)$ of algorithm 2

表 2　算法 1 和算法 2 在采样效率方面的比较

Table 2　Comparison between algorithm 1 and algorithm 2 in sampling efficiency

算法	算法 1	算法 2
采样数量 N_s	10	10
单条轨迹仿真时间 T_t (s)	10	10
采样周期 T_s (s)	0.001	0.001
总仿真时间 (s)	$N_s T_t = 100$	$N_s T_t \frac{2T_t}{T_s} = 2\,000\,000$

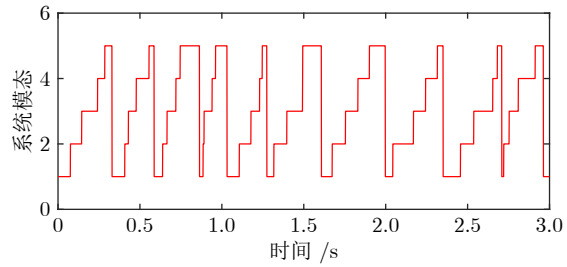


图 5　系统模态演化

Fig. 5　System mode evolution

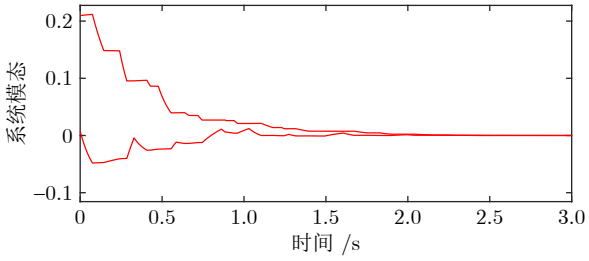


图 6　系统状态响应

Fig. 6　System state response

其中所有状态最终收敛至零. 结果验证了所提出的两种无模型策略梯度强化学习算法, 在求解具有部分或完全未知动力学的马尔科夫随机系统 (1) 的最优控制问题时的可行性和有效性.

4 结束语

本文针对具有部分或完全未知动力学的马尔科夫随机系统, 设计了部分无模型和完全无模型的基于策略梯度的强化学习算法, 以解决其最优控制问题. 首先, 通过获取策略梯度的解析形式, 开发了一种部分无模型的策略梯度强化学习算法, 从而提高了学习过程中的采样效率, 解决了部分未知动力学马尔科夫随机系统的最优控制问题. 然后, 通过对反馈增益的随机摄动估计策略梯度, 设计了一种完全无模型的策略梯度强化学习算法, 该算法不依赖马尔科夫随机系统的任何模型信息, 解决了完全未知动力学马尔科夫随机系统的最优控制问题. 通过数值仿真验证了所提出策略梯度强化学习算法的有效性和可行性. 在未来的研究中, 我们将进一步探索具有更高采样效率的完全无模型策略梯度强化学习算法, 并致力于解决控制增益初始约束可行性的问题.

References

1 Wang Ding, Wang Jiang-Yu, Qiao Jun-Fei. Data-driven policy optimization for stochastic systems involving adaptive critic. *Acta Automatica Sinica*, 2024, **50**(5): 980−990
(王鼎, 王将宇, 乔俊飞. 融合自适应评判的随机系统数据驱动策略优化. 自动化学报, 2024, **50**(5): 980−990)

2 Song Xiu-Lan, Li Yang-Yang, He De-Feng. Secure H_∞ platooning control for connected vehicles subject to external disturbance and random DoS attacks. *Acta Automatica Sinica*, 2024, **50**(2): 348−355
(宋秀兰, 李洋阳, 何德峰. 外部干扰和随机 DoS 攻击下的网联车安全 H_∞ 队列控制. 自动化学报, 2024, **50**(2): 348−355)

3 Kuppusamy S, Joo Y H, Kim H S. Asynchronous control for discrete-time hidden Markov jump power systems. *IEEE Transactions on Cybernetics*, 2022, **52**(9): 9943−9948

4 Man Jing-Tao, Zeng Zhi-Gang, Sheng Yin, Lai Jin-Gang. Finite-time formation of large-scale multi-agent systems based on an ODE-PDE approach. *Acta Automatica Sinica*, 2025, **51**(3): 631−642
(满景涛, 曾志刚, 盛银, 来金钢. 基于 ODE-PDE 的大规模多智能体系统有限时间编队. 自动化学报, 2025, **51**(3): 631−642)

5 Cui R H, Xie X J. Finite-time stochastic integral input-to-state stability and its applications. *Automatica*, 2023, **158**: Article No. 111311

6 Wang Le, Cui Kai, Jiang Xiu-Shan, Zhao Dong-Ya, Zhang Wei-Hai. Pareto optimal control of linear Markov jump stochastic systems. *Control Theory & Applications*, 2025, **42**(1): 59−66
(王乐, 崔凯, 蒋秀珊, 赵东亚, 张维海. 线性 Markov 跳变随机系统的 Pareto 最优控制. 控制理论与应用, 2025, **42**(1): 59−66)

7 Li F B, Du C L, Yang C H, Gui W H. Passivity-based asynchronous sliding mode control for delayed singular Markovian jump systems. *IEEE Transactions on Automatic Control*, 2018, **63**(8): 2715−2721

8 Li Gui-Lin, Wang Chuan-Rui, Ji Hai-Bo. Inverse optimal gain assignment control for a class of stochastic nonlinear systems with Markovian jump parameters. *Acta Automatica Sinica*, 2014, **40**(7): 1285−1294
(李桂林, 王传锐, 季海波. 具有 Markov 跳跃参数的一类随机非线性系统逆最优增益设计. 自动化学报, 2014, **40**(7): 1285−1294)

9 Liu Yue, Zhou Ping. Recent status and progress in optimal control of Markov jump linear systems. *Information and Control*, 2022, **51**(1): 54−68
(刘越, 周平. 马尔可夫跳变线性系统最优控制的研究现状与进展. 信息与控制, 2022, **51**(1): 54−68)

10 Xiong J L, Lam J, Gao H J, Ho D W C. On robust stabilization of Markovian jump systems with uncertain switching probabilities. *Automatica*, 2005, **41**(5): 897−903

11 Zhang Y, He Y, Wu M, Zhang J. Stabilization for Markovian jump systems with partial information on transition probability based on free-connection weighting matrices. *Automatica*, 2011, **47**(1): 79−84

12 Zhang K, Su R, Zhang H G. A novel resilient control scheme for a class of Markovian jump systems with partially unknown information. *IEEE Transactions on Cybernetics*, 2022, **52**(8): 8191−8200

13 Tian E G, Yue D, Wei G L. Robust control for Markovian jump systems with partially known transition probabilities and nonlinearities. *Journal of the Franklin Institute*, 2013, **350**(8): 2069−2083

14 Qi W H, Hou Y K, Park J H, Zong G D, Cao J D, Cheng J. SMC for discrete-time networked semi-Markovian switching systems with random DoS attacks and applications. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2023, **53**(3): 1982−1993

15 Wang Guo-Qing, Yao Feng-Qi. Finite-time stability of impulsive stochastic functional differential systems with bounded delays. *Control Theory & Applications*, 2023, **40**(9): 1569−1575
(王国庆, 姚凤麒. 具有有界时滞的脉冲随机泛函微分系统的有限时间稳定. 控制理论与应用, 2023, **40**(9): 1569−1575)

16 Wei Y L, Park J H, Qiu J B, Wu L G, Jung H Y. Sliding mode control for semi-Markovian jump systems via output feedback. *Automatica*, 2017, **81**: 133−141

17 Wu L G, Ho D W C. Sliding mode control of singular stochastic hybrid systems. *Automatica*, 2010, **46**(4): 779−783

18 Liu M, Zhang L X, Shi P, Zhao Y X. Sliding mode control of continuous-time Markovian jump systems with digital data transmission. *Automatica*, 2017, **80**: 200−209

19 Li F B, Du C L, Yang C H, Wu L G, Gui W H. Finite-time asynchronous sliding mode control for Markovian jump systems. *Automatica*, 2019, **109**: Article No. 108503

20 Du C L, Yang C H, Li F B, Gui W H. A novel asynchronous control for artificial delayed Markovian jump systems via output feedback sliding mode approach. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2019, **49**(2): 364−374

21 Li Z, Zhang Y, Wu A G. An inversion-free iterative algorithm for Riccati matrix equations in discrete-time Markov jump systems. *IEEE Transactions on Automatic Control*, 2022, **67**(9): 4754−4761

22 Dong S L, Liu L, Feng G, Liu M Q, Wu Z G, Zheng R H. Cooperative output regulation quadratic control for discrete-time heterogeneous multiagent Markov jump systems. *IEEE Transactions on Cybernetics*, 2022, **52**(9): 9882−9892

23 Jansch-Porto J P, Hu B, Dullerud G E. Policy optimization for Markovian jump linear quadratic control: Gradient method and global convergence. *IEEE Transactions on Automatic Control*, 2023, **68**(4): 2475−2482

24 Jansch-Porto J P, Hu B, Dullerud G E. Convergence guarantees of policy optimization methods for Markovian jump linear systems. In: Proceedings of the American Control Conference (ACC). Denver, USA: IEEE, 2020. 2882−2887

25 Chávez-Fuentes J R, Costa E F, Terra M H, Rocha K D T. The linear quadratic optimal control problem for discrete-time

Markov jump linear singular systems. *Automatica*, 2021, **127**: Article No. 109506

26 Lai J, Xiong J L, Shu Z. Model-free optimal control of discrete-time systems with additive and multiplicative noises. *Automatica*, 2023, **147**: Article No. 110685

27 Mohammadi H, Zare A, Soltanolkotabi M, Jovanović M R. Convergence and sample complexity of gradient methods for the model-free linear-quadratic regulator problem. *IEEE Transactions on Automatic Control*, 2022, **67**(5): 2435–2450

28 Bian T, Jiang Z P. Value iteration and adaptive dynamic programming for data-driven adaptive optimal control design. *Automatica*, 2016, **71**: 348–360

29 Li Yi-Chun, Liu Ze-Jiao, Hong Yi-Tian, Wang Ji-Chao, Wang Jian-Rui, Li Yi, et al. Multi-agent reinforcement learning based game: A survey. *Acta Automatica Sinica*, 2025, **51**(3): 540–558 (李艺春, 刘泽娇, 洪艺天, 王继超, 王健瑞, 李毅, 等. 基于多智能体强化学习的博弈综述. 自动化学报, 2025, **51**(3): 540–558)

30 Yuan Lei, Zhang Zi-Qian, Li Li-He, Guan Cong, Yu Yang. Progress on cooperative multi-agent reinforcement learning in open environment. *Scientia Sinica Informationis*, 2025, **55**(2): 217–268 (袁雷, 张子谦, 李立和, 管聪, 俞扬. 开放环境下的协作多智能体强化学习进展. 中国科学: 信息科学, 2025, **55**(2): 217–268)

31 Wang J, Zhao W, Cao J D, Park J H, Shen H. Reinforcement learning-based predefined-time tracking control for nonlinear systems under identifier-critic-actor structure. *IEEE Transactions on Cybernetics*, 2024, **54**(11): 6345–6357

32 Zhang Z Y, Mo Z B, Chen Y T, Huang J. Reinforcement learning behavioral control for nonlinear autonomous system. *IEEE/CAA Journal of Automatica Sinica*, 2022, **9**(9): 1561–1573

33 Gravell B, Esfahani P M, Summers T. Learning optimal controllers for linear systems with multiplicative noise via policy gradient. *IEEE Transactions on Automatic Control*, 2021, **66**(11): 5283–5298

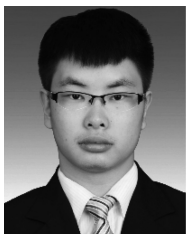
34 Ji Y D, Chizeck H J. Controllability, stabilizability, and continuous-time Markovian jump linear quadratic control. *IEEE Transactions on Automatic Control*, 1990, **35**(7): 777–788

35 Qian Y Y, Pang W J. An implicit sequential algorithm for solving coupled Lyapunov equations of continuous-time Markovian jump systems. *Automatica*, 2015, **60**: 245–250

36 Konda V R, Tsitsiklis J N. Actor-critic algorithms. In: Proceedings of the 13th International Conference on Neural Information Processing Systems (NIPS). Denver, USA: MIT Press, 1999. 1008–1014

37 Gao W N, Mynuddin M, Wunsch D C, Jiang Z P. Reinforcement learning-based cooperative optimal output regulation via distributed adaptive internal model. *IEEE Transactions on Neural Networks and Learning Systems*, 2022, **33**(10): 5229–5240

38 Borno I. Parallel computation of the solutions of coupled algebraic Lyapunov equations. *Automatica*, 1995, **31**(9): 1345–1347



杜城龙 中南大学自动化学院副教授. 2016 年获得哈尔滨工业大学学士学位, 2022 年获得中南大学博士学位. 主要研究方向为多智能体强化学习, 分布式控制与优化, 能源交通一体化和低空飞行器.
E-mail: chenglong_du@csu.edu.cn
(**DU Cheng-Long** Associate professor at the School of Automation, Central South University. He received his bachelor degree from Harbin Institute of Technology in 2016, and his Ph.D. degree from Central South University in 2022. His research interest covers multi-agent reinforcement learning, distributed control and optimization, energy and transportation integration, and low-altitude aircraft.)

韩 洁 中南大学自动化学院副教授. 2020 年获得中南大学博士学位. 主要研究方向为复杂工业过程智能建模与优化, 工业大模型和智能决策方法.
E-mail: hanjie@csu.edu.cn
(**HAN Jie** Associate professor at the School of Automation, Central South University. She received her Ph.D. degree from Central South University in 2020. Her research interest covers intelligent modeling and optimization of complex industrial process, industrial large-scale models, and intelligent decision-making methods.)

李繁飙 中南大学自动化学院教授. 2015 年获得哈尔滨工业大学博士学位. 主要研究方向为复杂工业过程智能控制与优化, 空天飞行器智能控制. 本文通信作者.
E-mail: fanbiaoli@csu.edu.cn
(**LI Fan-Biao** Professor at the School of Automation, Central South University. He received his Ph.D. degree from Harbin Institute of Technology in 2015. His research interest covers intelligent control and optimization of complex industrial process, and intelligent control of aerospace vehicle. Corresponding author of this paper.)

桂卫华 中国工程院院士, 中南大学自动化学院教授. 主要研究方向为复杂工业过程建模、优化与控制应用.
E-mail: gwh@csu.edu.cn
(**GUI Wei-Hua** Academician at Chinese Academy at Engineering, professor at the School of Automation, Central South University. His research interest covers modeling, optimization and control applications of complex industrial process.)