融合自适应评判的随机系统数据驱动策略优化

王 鼎 [1, 2, 3, 4] 王将宇 [1, 2, 3, 4] 乔俊飞 [1, 2, 3, 4]

摘 要 自适应评判技术已经广泛应用于求解复杂非线性系统的最优控制问题, 但利用其求解离散时间非线性随机系统的无限时域最优控制问题还存在一定局限性. 本文融合自适应评判技术, 建立一种数据驱动的离散随机系统折扣最优调节方法. 首先, 针对宽松假设下的非线性随机系统, 研究带有折扣因子的无限时域最优控制问题. 所提的随机系统 Q-learning 算法能够将初始的容许策略单调不增地优化至最优策略. 基于数据驱动思想, 随机系统 Q-learning 算法在不建立模型的情况下直接利用数据进行策略优化. 其次, 利用执行−评判神经网络方案, 实现了随机系统 Q-learning 算法. 最后, 通过两个基准系统, 验证本文提出的随机系统 Q-learning 算法的有效性.

关键词 自适应评判设计, 数据驱动, 离散系统, 神经网络, Q-learning, 随机最优控制

引用格式 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

DOI 10.16383/j.aas.c230678

Data-driven Policy Optimization for Stochastic Systems Involving Adaptive Critic

WANG Ding[1, 2, 3, 4] WANG Jiang-Yu[1, 2, 3, 4] QIAO Jun-Fei[1, 2, 3, 4]

Abstract Adaptive critic technology has been widely employed to solve the optimal control problems of complicated nonlinear systems, but there are some limitations to solve the infinite-horizon optimal problems of discrete-time nonlinear stochastic systems. In this paper, we establish a data-driven discounted optimal regulation method for discrete-time stochastic systems involving adaptive critic technology. First, we investigate the infinite-horizon optimal problems with the discount factor for stochastic systems under the relaxed assumption. The developed stochastic Q-learning algorithm can optimize an initial admissible policy to the optimal one in a monotonically nonincreasing way. Based on the data-driven idea, the policy optimization of the stochastic Q-learning algorithm is executed without a dynamic model. Then, the stochastic Q-learning algorithm is implemented by utilizing the actor-critic neural networks. Finally, two nonlinear benchmarks are given to demonstrate the overall performance of the developed stochastic Q-learning algorithm.

Key words Adaptive critic design, data-driven, discrete-time systems, neural networks, Q-learning, stochastic optimal control

Citation 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

现代工程与社会生活中广泛存在寻找最优方案的现实问题. 针对控制成本最小化问题, 最优控制提供了一系列解决方案. 对于一般的确定性非线性系统, 常见的技术手段是通过求解 Hamilton-Jacobi-Bellman 方程获得最优策略[1]. 这一方程通常包含难以直接求解的微分方程. 动态规划为最优控制问题提供了一种简化的求解方法, 即将一个多级决策问题转化为多个单级决策问题[2]. 然而, 在处理复杂高维问题时, 动态规划往往会面临“维数灾”问题[3]. 因此, 基于自学习特性, 提出自适应评判 (或自适应动态规划) 技术[4−6] 并用以解决复杂非线性系统的最优控制问题. 尽管强化学习与自适应评判在不同方面强调了各自的特点, 但它们都试图通过学习的方式不断靠近最优目标. 在过去十年里, 涌现了大量基于强化学习的自适应评判技术, 例如, 平行控制[7−8]、演化学习控制[9]、事件触发控制[10−12]、智能工业控制[13−14] 等. 在某种程度上, 强化学习极大启发了智能控制领域的创新. 结合目前的发展形势, 本

收稿日期 2023-11-02 录用日期 2024-01-08
Manuscript received November 2, 2023; accepted January 8, 2024
国家自然科学基金 (62222301, 61890930-5, 62021003), 科技创新 2030 ——“新一代人工智能”重大项目 (2021ZD0112302, 2021ZD0112301) 资助
Supported by National Natural Science Foundation of China (62222301, 61890930-5, 62021003) and National Key Research and Development Program of China (2021ZD0112302, 2021ZD0112301)
本文责任编委 杨涛
Recommended by Associate Editor YANG Tao
1. 北京工业大学信息学部 北京 100124 2. 计算智能与智能系统北京市重点实验室 北京 100124 3. 北京人工智能研究院 北京 100124 4. 智慧环保北京实验室 北京 100124
1. Faculty of Information Technology, Beijing University of Technology, Beijing 100124 2. Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing 100124 3. Beijing Institute of Artificial Intelligence, Beijing 100124 4. Beijing Laboratory of Smart Environmental Protection, Beijing 100124

文将融合自适应评判技术实现一种自学习的策略优化算法.

近年来, 许多新兴技术的重要特点是拥有大量的数据信息, 为数据驱动算法设计提供了现实基础. 2022 年, OpenAI 研发了一款现象级的聊天机器人程序 —— ChatGPT[15]. 由人工智能技术驱动, ChatGPT 模型通过连接大量的语言库来迭代学习. 在自适应评判领域, 数据驱动思想同样取得了丰硕的研究成果. Luo 等[16] 提出了一种数据驱动的策略梯度自适应动态规划算法, 以解决离散系统无模型最优控制问题, 并给出学习过程中策略的收敛性分析. 基于策略迭代 Q-learning 结构, 提出一种数据驱动的无模型算法[17] 并用于求解线性系统博弈问题, 且通过理论证明了迭代 Q 函数的最优性. Lin 等[18] 基于经验回放提出了策略梯度自适应评判方法, 实现了离散时间非线性系统的无模型最优跟踪控制. 文献 [19] 中, 报道了一种基于柔和执行–评判的数据驱动学习方法用以解决列车节能控制问题. 可见, 在大数据迅速发展的背景下, 数据驱动是实现智能控制的有效方案. 然而, 上述研究的主要对象是确定性系统, 在一定程度上限制了自学习算法的适用范围. 由于机械振动和测量误差等因素的存在, 系统动力学中可能会引入随机扰动. 尽管扰动会增加方案设计和理论分析的难度, 但含随机项的控制器设计更具有实际价值.

鲁棒控制一般会考虑最坏的扰动情况, 然而, 对于某些具有统计特性的扰动而言, 这可能会导致过于保守的设计. 相较于鲁棒控制, 随机控制的主要优势体现在处理具有统计特性噪声时丰富的数学工具, 使得控制器设计更加灵活. 针对随机扰动, 已经有一些学者尝试利用自适应评判技术求解随机系统的最优控制问题. Wei 等[20] 提出了一种解决连续时间最优控制的问题自适应评判方法, 该算法将扰动转换为确定性系统零和博弈问题, 但需要建立系统模型和设计相应的博弈代价函数. Liang 等[21] 提出了一种改进的随机系统值迭代算法, 用于求解离散时间随机系统最优控制问题. Pang 等[22] 研究了带有多种随机扰动的线性系统最优控制问题, 利用自适应评判技术实现了最优策略的脱策学习. 最近, 文献 [23] 提出一种基于模型的连续随机系统策略迭代算法, 通过一个初始容许策略将代价函数的期望最小化求解最优控制策略. 目前, 自适应评判领域针对离散时间非线性随机系统的研究还存在局限性, 尤其是关于无模型数据驱动的相关算法和理论.

为了保证随机系统最优控制的可行性, 通常需要引入容许控制的概念确保代价函数有界. 确定性系统相对容易确保无穷时域代价函数存在上界. 对于常见二次型效用函数, 确定性系统利用镇定策略使得状态到达平衡点后效用函数便不再累加, 从而确保无限时域代价函数有上界. 然而, 随机系统可能在平衡点附近按一定概率运动, 导致直接定义的无限时域代价函数无上界. 对此, 关于随机系统的最优控制研究一般假设系统的随机扰动在到达平衡点后消失[24−25]. 本文力求在研究随机系统无限时域最优控制问题的基础上, 放宽对随机扰动形式的假设, 以扩大本文所提方法的应用范围. 具体而言, 通过定义带有折扣因子[26−27] 的代价函数, 放松对系统严格的假设且仍能研究无限时域最优控制问题.

Lincoln 和 Rantzer[28] 首先提出松弛动态规划用于解决最优控制问题中的“维数灾”, 其核心是利用折扣因子适当放松最优性以减少算法实现成本. 文献 [28] 中设计折扣因子保证库存订单控制代价函数有界, 从而获得次优策略. Ha 等[29] 分析了非线性系统稳定性与折扣因子的关系, 结果表明如果折扣因子选取不当可能导致折扣最优策略不是镇定策略. 此外, 基于自适应评判技术的折扣最优控制问题衍生出许多分支, 如跟踪控制问题[30] 和非零平衡问题[31] 等. 基于上述研究[28−31], 可知对于一般的最优控制问题非必要可不引入折扣因子. 这是由于引入折扣会损失代价函数的最优性[28], 且对系统的稳定性会带来负面影响[29]. 然而, 对于一些无法直接评估无穷时域代价函数的问题可通过折扣因子保证代价函数有界. 需要指出的是在严格的最优性定义下, 自适应评判技术将不再适用于本文的研究问题. 基于此, 本文对折扣因子的应用进行了合理推广, 同时使得自适应评判技术的学习优势在更复杂情况下得以验证.

综上, 本文提出了一种基于无模型 Q-learning 技术的数据驱动策略优化算法, 以解决离散时间随机系统的折扣无限时域最优控制问题. 在折扣因子的作用下, 本文所提方法放宽了对于随机系统扰动形式的假设, 拓宽了自适应评判技术的应用范围. 通过由任意容许策略初始化的 Q 函数, 我们证明了所提算法能够使得迭代 Q 函数序列单调不增地迭代至最优. 此外, 本文所提方法可以直接从实际系统获取的数据中学习, 减少了随机系统建模的负担. 最后, 通过两个基准系统测试了本文所提算法的综合性能.

在本文中, $\mathbf{R}$ 和 $\mathbf{N}$ 分别代表所有实数集合和非负整数集合. $\mathbf{R}^n$ 和 $\|\mathbf{R}^n\|$ 分别代表由 n 维实向量组成的欧氏空间和对应的范数. $\mathbf{R}^{n\times m}$ 代表 $n\times m$ 实矩阵组成的空间. Ω 代表 $\mathbf{R}^n$ 上一个紧集. I_n 代表 $n\times n$ 维的单位矩阵. $\mathrm{E}\{\cdot\}$ 代表随机变量的数学期望. T 代表转置运算.

1 问题描述

考虑一类离散非线性随机系统

$$x_{k+1} = F(x_k, u_k, \omega_k), k \in \mathbf{N} \tag{1}$$

其中, $x_k \in \Omega_x \subset \mathbf{R}^n$, $u_k \in \Omega_u \subset \mathbf{R}^m$ 和$\omega_k \in \mathbf{R}^l$ 分别代表系统的状态向量, 控制输入以及随机扰动. 便于后续分析, 本文所得结论基于如下假设[25]:

假设 1. 系统函数 $F(x, u, 0)$ 在包含原点的集合上 Lipschitz 连续且有 $F(0, 0, 0) = 0$. 将 $x = 0$ 称作在控制 $u = 0$ 和无扰动作用下的一个平衡点.

假设 2. 随机扰动 ω 是时不变的, 且仅与当前系统状态与控制输入有关. 同时存在一个连续的策略 $u(x)$ 使得系统能够从初始状态转移到平衡点的邻域内, 即 $\mathrm{E}\{F(0, 0, \omega)\} = 0$.

本文研究带有折扣因子的无限时域随机系统最优控制问题. 控制器的设计目标是对初始策略 $u(x)$: $\mathbf{R}^n \to \mathbf{R}^m$, 进行策略优化得到新的策略 $\mu(x)$, 能够将系统从初始状态 $x_0 \in \Omega_x$ 镇定到平衡点的邻域内, 同时使得无限时域的代价函数

$$J_\mu(x_0) = \mathrm{E}\left\{\sum_{p=0}^{\infty} \lambda^p U(x_p, \mu(x_p))\right\} \tag{2}$$

在所有的可行策略中达到最小, 其中, $\lambda \in (0, 1)$是折扣因子, $U(x_p, \mu(x_p)) \geq 0$ 是效用函数, 且 $U(0, 0) = 0$. 由于折扣因子 λ 具有"遗忘"特性, 因此越远离当前状态的效用比重越小, 即随机系统的无限时域最优控制问题是近似的. 在不产生歧义的情况下, $J_\mu(x_k)$简写为 $J(x_k)$ 或 $J(x)$. 对于最优控制问题 (2), 目标策略不仅能够在 Ω_x 镇定随机系统, 同时应使得对应的代价函数有界, 这对应着容许控制策略的概念. 区别于确定性系统, 定义随机系统中的容许控制策略如下[23]:

定义 1. 如果策略 $\mu(x)$ 对所有的 $x \in \Omega_x$ 是连续的, $\mu(0) = 0$, $\mu(x)$ 能够镇定随机系统 (1), 且使得代价函数 (2) 有界, 则将策略 $\mu(x)$ 视为可容许控制族 $\mathcal{A}(\Omega_x)$ 中的一个容许控制, 记作 $\mu(x) \in \mathcal{A}(\Omega_x)$.

根据最优控制理论, 最优代价函数满足

$$J^*(x_0) = \min_{\mu(\cdot)} \mathrm{E}\left\{\sum_{p=0}^{\infty} \lambda^p U(x_p, \mu(x_p))\right\} \tag{3}$$

由 Bellman 最优性原理, 对所有的 $x \in \Omega_x$, 离散随机系统的最优代价函数满足如下 Bellman 最优方程:

$$J^*(x) = \min_{\mu(\cdot)} \mathrm{E}\{U(x, \mu(x)) + \lambda J^*(F(x, \mu(x), \omega))\} \tag{4}$$

相应的最优策略为

$$\mu^*(x) = \arg\min_{\mu(\cdot)} \mathrm{E}\{U(x, \mu(x)) + \lambda J^*(F(x, \mu(x), \omega))\} \tag{5}$$

本文的目标是, 基于数据驱动技术, 即在系统动态未知的情况下仅通过测量系统的状态向量 x 和控制输入 u, 在代价函数 (2) 学习随机非线性系统的近似最优策略.

注 1. 关于如何求解式 (5) 中的最优策略 $\mu^*(x)$, 需关注如下难题: 求解最优策略时, 需要得到最优代价函数 $J^*(x)$ 和系统的全部动态信息 $F(x, \mu(x), \omega)$. 然而, 对于非线性随机系统, 通常难以直接获得最优代价函数 $J^*(x)$ 的解析解. 此外, 关于系统动态信息的要求对非线性随机系统而言也是一项挑战, 且与本文数据驱动的设计目标矛盾. 其次, 区别于确定系统的控制序列, 由于需要应对突发状态, 最优策略更具应用价值的同时更难获取. 获取难度主要体现在求解控制序列只需针对目标状态求解, 而策略的获取则需要更多的数据来形成一个完整的策略. 因此, 在上述难题下, 需要考虑通过学习的途径获取近似最优策略以满足随机系统的控制需求.

2 随机系统策略优化方案

本节我们主要关注随机系统 Q-learning 算法的设计, 包括对于初始策略的评估和策略优化方案的建立. 此外, 分析了随机系统 Q-learning 算法的收敛性和单调性, 即证明本文算法能够将一个容许的初始策略迭代调优, 使得代价函数单调不增地收敛到最优 Q 函数.

2.1 随机系统 Q-learning 算法设计

根据代价函数的定义式 (2), 这样的评估仅关注当前状态的代价. 为了实现数据驱动技术, 基于文献 [16], 本文引入 Q 函数评估状态–控制对的综合代价, 将其定义为

$$Q(x, a) = \mathrm{E}\{U(x, a) + \lambda J_\mu(F(x, a, \omega))\} \tag{6}$$

其中, $U(x, a)$ 表示行为控制 $a \in \Omega_u$ 作用下的效用函数, $J_\mu(F(x, a, \omega))$ 表示后续状态 $F(x, a, \omega)$ 下代价函数的值. 注意, 根据定义式 (2), 代价函数 $J_\mu(x)$ 对应着一个容许的策略 $\mu(x)$. 因此, Q 函数意味着基于策略 $\mu(x)$, 评估状态 x 对应的行为控制 a 的综合代价. 基于此, 最优 Q 函数可表示为

$$Q^*(x, a) = \mathrm{E}\{U(x, a) + \lambda J^*(F(x, a, \omega))\} \tag{7}$$

对应的最优策略为

$$\mu^*(x) = \arg\min_a Q^*(x, a) \tag{8}$$

综合式 (5) 和 (8), 寻找最优策略的问题转换为寻找最优的 Q 函数, 从而避免对于系统动态信息的要求.

接下来, 我们介绍本文的数据驱动策略优化方案. 对于所有的$x_k \in \Omega_x$和给定的初始容许策略 $\mu(x) \in \mathcal{A}(\Omega_x)$, 可以通过求解如下方程获得初始 Q 函数

$$Q^{(0)}(x_k, a) = \mathrm{E}\{U(x_k, a) + \lambda Q^{(0)}(x_{k+1}, \mu(x_{k+1}))\} \tag{9}$$

其中, $x_{k+1} = F(x_k, a, \omega_k)$.

基于迭代自适应评判方法的基本思路, 本文构建两个序列不断优化初始策略, 即 Q 函数序列 $\{Q^{(\rho)}(x, a)\}$ 和迭代策略序列 $\{\mu^{(\rho)}(x)\}$, 其中 $\rho \in \mathbf{N}$ 代表迭代指标. 通过策略评估与策略提升不断更新 Q 函数序列和迭代策略序列, 直到收敛到近似最优策略. 具体而言, 本文所提出的随机系统 Q-learning 算法在如下两个过程间往复迭代, 即策略提升

$$\mu^{(\rho)}(x_k) = \arg\min_a Q^{(\rho)}(x_k, a) \tag{10}$$

和策略评估

$$Q^{(\rho+1)}(x_k, a) = \mathrm{E}\big\{U(x_k, a) + \lambda Q^{(\rho)}(x_{k+1}, \mu^{(\rho)}(x_{k+1}))\big\} \tag{11}$$

注 2. 关于随机系统数据驱动策略优化算法有以下几点说明: 随机系统 Q-learning 算法的实现过程可完全不依赖系统的动态信息. 通过观察策略评估式 (11), 系统在 $k+1$ 时刻的状态信息可通过当前状态 x_k 和行为控制 a 提前产生, 并且这些三元组信息 (x_k, a, x_{k+1}) 不会在迭代的过程中发生改变. 注意, 行为控制 a 是由探索随机系统的行为策略产生, 而 $\mu^{(\rho)}(x)$ 是本文的目标策略. 本文的目标策略与行为策略是不同的, 不需要每次使用当前目标策略收集数据, 而是通过引入行为策略探索环境, 提高数据利用率. 具体而言, 本文的随机系统 Q-learning 是一种脱策学习算法, 能够有效避免在线优化过程中探索不足导致的策略局部最优. 最后, 策略优化时, 仅利用系统的历史数据可实现一个已知策略的调优, 而无需重新学习策略或收集信息. 上述特点为数据驱动策略优化算法提供强有力的应用支撑.

2.2　算法收敛性分析

接下来分析随机系统 Q-learning 算法的收敛性和单调性.

定理 1. 对于所有的 $x \in \Omega_x$ 和 $a \in \Omega_u$, 令 Q 函数序列 $\{Q^{(\rho)}(x, a)\}$和策略序列 $\{\mu^{(\rho)}(x)\}$ 分别由式 (11) 和 (10) 更新. 若初始 Q 函数 $Q^{(0)}(x, a)$ 由式 (9) 产生, 则迭代 Q 函数 $Q^{(\rho)}(x, a)$ 随迭代步 $\rho \to \infty$ 收敛到最优, 即

$$\lim_{\rho\to\infty} Q^{(\rho)}(x, a) = Q^*(x, a) \tag{12}$$

证明. 利用数学归纳法证明随机系统 Q-learning 算法的收敛性. 首先, 通过一个容许策略 $\mu(x)$ 对 Q 函数 $Q^{(0)}(x, a)$ 初始化. 根据最优 Q 函数的定义, 可得

$$\begin{aligned} Q^*(x, a) = {} & \mathrm{E}\{U(x, a) + \lambda J^*(F(x, a, \omega))\} \le \\ & \mathrm{E}\{U(x, a) + \lambda J(F(x, a, \omega))\} = \\ & Q^{(0)}(x, a) \end{aligned} \tag{13}$$

由不等式 (13) 可知, 当 $\rho = 0$, $Q^*(x, a) \le Q^{(\rho)}(x, a)$ 成立. 假设对于 ρ 我们有 $Q^*(x, a) \le Q^{(\rho)}(x, a)$ 成立, 则

$$\begin{aligned} & Q^{(\rho+1)}(x_k, a) = \\ & \quad \mathrm{E}\big\{U(x_k, a) + \lambda Q^{(\rho)}(x_{k+1}, \mu^{(\rho)}(x_{k+1}))\big\} \ge \\ & \quad \mathrm{E}\big\{U(x_k, a) + \lambda Q^*(x_{k+1}, \mu^{(\rho)}(x_{k+1}))\big\} \ge \\ & \quad \mathrm{E}\big\{U(x_k, a) + \min_{u(\cdot)} \lambda Q^*(x_{k+1}, u(x_{k+1}))\big\} = \\ & \quad \mathrm{E}\big\{U(x_k, a) + \lambda J^*(x_{k+1})\big\} = \\ & \quad Q^*(x_k, a) \end{aligned} \tag{14}$$

因此, 由数学归纳法, 对于所有的 $\rho \in \mathbf{N}$, 我们有 $Q^*(x, a) \le Q^{(\rho)}(x, a)$ 成立. 同理,

$$\lim_{\rho\to\infty} Q^{(\rho)}(x, a) \ge Q^*(x, a) \tag{15}$$

考虑任意一个镇定策略 $\breve{\mu}(x)$. 当作用策略 $\breve{\mu}(x)$ 到随机非线性系统式 (1) 时, 系统状态满足 $\lim_{k\to\infty} \mathrm{E}\{x_k\} = 0$. 接下来根据随机系统 Q-learning 算法的迭代过程式 (10) 和 (11), 可得

$$\begin{aligned} & Q^{(\rho)}(x_k, a) = \\ & \quad \mathrm{E}\big\{U(x_k, a) + \lambda Q^{(\rho-1)}(x_{k+1}, \mu^{(\rho-1)}(x_{k+1}))\big\} = \\ & \quad \mathrm{E}\big\{U(x_k, a) + \min_{u(\cdot)} \lambda Q^{(\rho-1)}(x_{k+1}, u(x_{k+1}))\big\} \le \\ & \quad \mathrm{E}\big\{U(x_k, a) + \lambda Q^{(\rho-1)}(x_{k+1}, \breve{\mu}(x_{k+1}))\big\} = \\ & \quad \mathrm{E}\big\{U(x_k, a) + \lambda U(x_{k+1}, \breve{\mu}(x_{k+1})) + \\ & \quad \lambda^2 Q^{(\rho-2)}(x_{k+2}, \mu^{(\rho-2)}(x_{k+2}))\big\} \le \\ & \qquad \vdots \\ & \quad \mathrm{E}\left\{U(x_k, a) + \sum_{p=1}^{\rho-1} \lambda^p U(x_{k+p}, \breve{\mu}(x_{k+p})) + \right. \\ & \quad \left. \lambda^\rho Q^{(0)}(x_{k+\rho}, \mu^{(0)}(x_{k+\rho}))\right\} \end{aligned} \tag{16}$$

考虑到 $Q^{(0)}(x, a)$ 由一个容许策略 $\mu(x)$ 初始化, 我们有

$$\lim_{k\to\infty} \mathrm{E}\left\{Q^{(0)}(x_k,\, \mu(x_k))\right\} = 0 \tag{17}$$

根据式 (17) 并对不等式 (16) 取极限, 可得

$$\begin{aligned}&\lim_{\rho\to\infty} Q^{(\rho)}(x_k,\, a) \le \\ &\quad \mathrm{E}\left\{U(x_k,\, a) + \sum_{p=1}^{\infty} \lambda^p U(x_{k+p},\, \breve{\mu}(x_{k+p}))\right\} = \\ &\quad \mathrm{E}\left\{U(x_k,\, a) + \lambda J_{\breve{\mu}}(x_{k+1})\right\} = \\ &\quad Q_{\breve{\mu}}(x_k,\, a)\end{aligned} \tag{18}$$

由不等式 (15) 和 (18), 易得

$$Q^*(x,\, a) \le \lim_{\rho\to\infty} Q^{(\rho)}(x,\, a) \le Q_{\breve{\mu}}(x,\, a) \tag{19}$$

若取 $\breve{\mu}(x) = \mu^*(x)$, 则不等式 (19) 可改写为

$$Q^*(x,\, a) \le \lim_{\rho\to\infty} Q^{(\rho)}(x,\, a) \le Q_{\mu^*}(x,\, a) = Q^*(x,\, a) \tag{20}$$

最后, 综上所述可得 $\lim_{\rho\to\infty} Q^{(\rho)}(x,\, a) = Q^*(x,\, a)$. □

定理 2. 对于所有的 $x \in \Omega_x$ 和 $a \in \Omega_u$, 令 Q 函数序列 $\{Q^{(\rho)}(x,\, a)\}$ 和策略序列 $\{\mu^{(\rho)}(x)\}$ 分别由式 (11) 和 (10) 更新. 若初始 Q 函数 $Q^{(0)}(x,\, a)$ 由式 (9) 产生, 则对于所有的 $\rho \in \mathbf{N}$, 迭代 Q 函数序列 $\{Q^{(\rho)}(x,\, a)\}$ 是单调不增的, 即

$$Q^{(\rho+1)}(x,\, a) \le Q^{(\rho)}(x,\, a) \tag{21}$$

证明. 首先, 由初始 Q 函数式 (9) 和策略评估式 (11), 可得

$$\begin{aligned}&Q^{(1)}(x_k,\, a) = \\ &\quad \mathrm{E}\left\{U(x_k,\, a) + \lambda Q^{(0)}(x_{k+1},\, \mu^{(0)}(x_{k+1}))\right\} = \\ &\quad \mathrm{E}\left\{U(x_k,\, a) + \min_{u(\cdot)} \lambda Q^{(0)}(x_{k+1},\, u(x_{k+1}))\right\} \le \\ &\quad \mathrm{E}\left\{U(x_k,\, a) + \lambda Q^{(0)}(x_{k+1},\, \mu(x_{k+1}))\right\} = \\ &\quad Q^{(0)}(x_k,\, a)\end{aligned} \tag{22}$$

显然, 根据不等式 (22), 结论 $Q^{(\rho+1)}(x,\, a) \le Q^{(\rho)}(x,\, a)$ 对 $\rho = 0$ 成立. 接下来, 假设不等式 (21) 对 $\rho - 1$ 成立, $\rho \in 1,\, 2,\, \cdots$, 我们有

$$\begin{aligned}&Q^{(\rho+1)}(x_k,\, a) = \\ &\quad \mathrm{E}\left\{U(x_k,\, a) + \lambda Q^{(\rho)}(x_{k+1},\, \mu^{(\rho)}(x_{k+1}))\right\} = \\ &\quad \mathrm{E}\left\{U(x_k,\, a) + \min_{u(\cdot)} \lambda Q^{(\rho)}(x_{k+1},\, u(x_{k+1}))\right\} \le \\ &\quad \mathrm{E}\left\{U(x_k,\, a) + \min_{u(\cdot)} \lambda Q^{(\rho-1)}(x_{k+1},\, u(x_{k+1}))\right\} = \\ &\quad Q^{(\rho)}(x_k,\, a)\end{aligned} \tag{23}$$

由数学归纳法可知, 对所有的 $\rho \in \mathbf{N}$ 迭代 Q 函数序列是单调不增的, 即 $Q^{(\rho+1)}(x,\, a) \le Q^{(\rho)}(x,\, a)$. □

注 3. 通过定理 1 证明, 由容许策略初始化的 Q 函数 $Q^{(0)}(x,\, a)$, 可以收敛到最优值. 结合定理 2 可知, 随机系统 Q-learning 算法的代价函数序列能单调不增地收敛到最优值. 事实上, 我们还可选取较小的折扣因子来保证算法的收敛性. 但折扣因子过小时会对系统的稳定性产生负面影响, 即使用小折扣策略可能会导致系统不稳定[29]. 因此, 本文通过一个稳定的折扣初始策略确保优化过程中系统的稳定性.

3 随机系统策略优化算法的数据驱动实现

本节给出随机系统 Q-learning 算法的数据驱动实现.

本文通过引入执行–评判结构实现随机系统 Q-learning 算法, 其中执行网络用于策略优化, 评判网络用于评估迭代策略. 在评判网络中, 将近似的 Q 函数定义为

$$\hat{Q}^{(\rho)}(x,\, a) = \Phi^{\mathrm{T}}(x,\, a)\vartheta_Q^{(\rho)} \tag{24}$$

其中, $\Phi(x,\, a) = [\phi_1(x,\, a),\, \phi_2(x,\, a),\, \cdots,\, \phi_{L_Q}(x,\, a)]^{\mathrm{T}}$ 代表评判网络的激活函数, $\vartheta_Q^{(\rho)} \in \mathbf{R}^{L_Q}$ 代表评判网络权值向量.

相应地, 在执行网络中, 将近似迭代策略 $\mu^{(\rho)}(x)$ 定义为

$$\hat{\mu}^{(\rho)}(x) = \left(\Upsilon^{\mathrm{T}}(x)\vartheta_\mu^{(\rho)}\right)^{\mathrm{T}} \tag{25}$$

其中, $\Upsilon(x) = [\upsilon_1(x),\, \upsilon_2(x),\, \cdots,\, \upsilon_{L_\mu}(x)]^{\mathrm{T}}$ 代表执行网络的激活函数, $\vartheta_\mu^{(\rho)} \in \mathbf{R}^{L_\mu \times m}$ 代表执行网络权值向量.

考虑策略提升式 (10), 通过求解优化问题, 执行网络的权值可更新为

$$\begin{aligned}\vartheta_\mu^{(\rho)} = &\ \arg\min_{\vartheta_\mu} \left\{\hat{Q}^{(\rho)}(x,\, \mu(x))\right\} = \\ &\arg\min_{\vartheta_\mu} \left\{\Phi^{\mathrm{T}}\left(x,\, \left(\Upsilon^{\mathrm{T}}(x)\vartheta_\mu\right)^{\mathrm{T}}\right)\vartheta_Q^{(\rho)}\right\}\end{aligned} \tag{26}$$

针对权值更新过程式 (26), 可以采用多种方式进行调优, 如梯度法、群智能方法等.

考虑策略评估式 (11), 评判网络的权值可更新为

$$\begin{aligned}\hat{Q}^{(\rho+1)}(x_k,\, a) = &\ \mathrm{E}\Big\{U(x_k,\, a) + \\ &\lambda\hat{Q}^{(\rho)}(x_{k+1},\, \hat{\mu}^{(\rho)}(x_{k+1}))\Big\}\end{aligned} \tag{27}$$

其中, 效用函数为 $U(x,\,a)=x^{\mathrm{T}}\mathcal{Q}x+a^{\mathrm{T}}\mathcal{R}a$, $\mathcal{Q}$ 和 $\mathcal{R}$ 是维度匹配的正定矩阵. 将式 (24) 代入近似策略评估式 (27), 可得

$$\Phi^{\mathrm{T}}(x_k,\,a)\vartheta_Q^{(\rho+1)}=\mathrm{E}\Big\{U(x_k,\,a)+\lambda\Phi^{\mathrm{T}}(x_{k+1},\,\hat{\mu}^{(\rho)}(x_{k+1}))\vartheta_Q^{(\rho)}\Big\}\tag{28}$$

利用不同探索行为 a, 通过与真实系统交互收集策略优化方案所需要的数据集, 并将其表示为

$$D_M=\big\{x_k^{[M]},\,a^{[M]},\,\bar{x}_{k+1}^{[M]}|M\in\mathbf{N}\big\}\tag{29}$$

其中, M 代表数据集的大小. $\bar{x}_{k+1}^{[M]}=\{x_{k+1}^{[M],\,[i]}|i=1,\,2,\,\cdots,\,I\}$ 代表由相同的状态 $x_k^{[M]}$ 和控制输入 $a^{[M]}$ 作用下的下一时刻随机系统状态, 其中 $I\in\mathbf{N}^+$. 当选取的重复次数 I 足够大时, 由大数定律可以估计下一时刻的状态期望, 将每一组数据的残差表示为

$$\varepsilon^{(\rho+1)}(x_k^{[M]})=\mathrm{E}\Big\{\Phi^{\mathrm{T}}(x_k^{[M]},\,a^{[M]})\vartheta_Q^{(\rho+1)}-U(x_k^{[M]},\,a^{[M]})-\lambda\Phi^{\mathrm{T}}(\bar{x}_{k+1}^{[M]},\,\hat{\mu}^{(\rho)}(\bar{x}_{k+1}^{[M]}))\vartheta_Q^{(\rho)}\Big\}\tag{30}$$

通过最小化数据集的残差和, 评判网络的权值由如下形式更新:

$$\vartheta_Q^{(\rho+1)}=\arg\min_{\vartheta_Q^{(\rho+1)}}\left\{\sum_{p=1}^{M}\varepsilon^{(\rho+1)}(x_k^{[p]})\right\}\tag{31}$$

观察评判网络的权值更新过程式 (30) 和 (31), 其主要任务是拟合一个新的近似 Q 函数. 因此, 各种神经网络训练方法均可实现神经网络参数调优, 例如最小二乘、多元线性回归、梯度法以及智能优化方法等. 最后, 基于 Q-learning 的随机系统策略优化由算法 1 给出.

算法 1. 基于 Q-learning 的随机系统策略优化

初始化. 选择效用函数矩阵 $\mathcal{Q}$ 和 $\mathcal{R}$. 收集数据集 D_M. 选择折扣因子 λ. 设置算法的停止误差 ϵ. 设置最大迭代次数 $\rho_{\max}$. 利用式 (9) 初始化 Q 函数 $Q^{(0)}(x,\,a)$, 并令 $\rho=0$.

数据驱动脱策学习:

步骤 1. 实现策略提升. 通过求解优化问题式 (26), 更新执行网络权值 $\vartheta_\mu^{(\rho)}$.

步骤 2. 实现策略评估. 通过求解函数逼近问题式 (31), 更新评判网络权值 $\vartheta_Q^{(\rho+1)}$.

步骤 3. 判断条件 $\|\vartheta_Q^{(\rho+1)}-\vartheta_Q^{(\rho)}\|\le\epsilon$. 若成立则转步骤 6, 否则转步骤 4.

步骤 4. 更新迭代指标, 令 $\rho=\rho+1$.

步骤 5. 判断条件 $\rho>\rho_{\max}$. 若成立则转步骤 6, 否则转步骤 1.

步骤 6. 输出执行网络权值 $\hat{\vartheta}_\mu^*$ 与评判网络权值 $\hat{\vartheta}_Q^*$.

注 4. 由近似策略评估式 (27), 学习过程中需要求解关于随机状态 x_{k+1} 的控制输入 $\hat{\mu}^{(\rho)}(x_{k+1})$. 在确定性系统中[32], 为了减少计算量和降低执行网络的累积误差, 通常的技术手段是仅更新迭代过程的控制序列而不更新执行网络权值, 并在算法收敛后训练一次执行网络. 由于随机系统可能出现突发状态, 导致控制序列失效. 这意味着无法忽视执行网络训练, 从而会带来额外的计算负担与神经网络累积误差. 因此, 对于随机系统而言, 高精度函数逼近工具的重要性尤其突出.

注 5. 算法 1 直接从数据集学习近似最优 Q 函数, 即迭代过程完全不依赖系统动态, 同时避免建模误差. 相较于建模数据驱动, 预学习过程往往会增加算法的学习负担. 此外, 对于模型误差的分析是一项棘手的工作, 文献 [33] 给出了相应的分析过程. 本文所提方法的优势主要体现在实现简单且不需要引入额外的辨识工具, 而理论分析主要体现在不需要考虑由于模型辨识带来的误差分析. 在学习过程中, 执行网络基于当前的评判网络的估计产生一个提升的策略. 之后, 评判网络对提升的策略进行一轮评估得到新的评判网络. 随着迭代的进行, 执行网络和评判网络最终会收敛到对应的最优值. 定理 1 表明迭代指标 $\rho\to\infty$ 时, 迭代 Q 函数收敛至最优. 实际算法实现过程则依靠较小的停止误差 ϵ 终止程序, 以获得近似最优策略.

4　基准实验分析

本节通过两个基准系统, 验证所提的随机系统 Q-learning 算法具有显著策略优化能力以及无模型数据驱动学习能力.

4.1　基准系统 I

考虑一个二阶非线性扭摆系统[34], 用于测试控制算法的综合性能. 本文采用带有随机噪声的扭摆系统, 其状态空间表达式为

$$x_{k+1}=\begin{bmatrix}x_k(1)+0.05x_k(2)\\-0.245\sin x_k(1)+0.99x_k(2)+0.05u_k\end{bmatrix}+\begin{bmatrix}\omega_k\\\omega_k\end{bmatrix}+\delta_k x_k\tag{32}$$

其中, $\omega_k\sim\mathrm{N}(0,\,0.01^2)$ 和 $\delta_k\sim\mathrm{N}(0,\,0.05^2)$ 是两个服从高斯分布的独立噪声. 由系统方程可知, ω_k 是独

立于状态和控制的噪声, 这意味着使用稳定策略只能调节系统状态到平衡点的一个邻域. 而噪声 δ_k 则是取决于系统状态的, 即状态越偏离平衡点时 δ_k 对系统的影响越显著. 对于随机系统式 (32), 控制器的设计目标是在容许控制族中寻找近似最优策略 $\hat{\mu}^*(x) \in \mathcal{A}(\Omega_x)$ 从而使得系统镇定.

依据算法 1, 从真实的系统中收集 1 331 组数据构成数据集 D_M, 其中 $\Omega_x = \{|x(1)| \leq 1, |x(2)| \leq 1\}$, $\Omega_\mu = \{|u| \leq 2\}$. 样本选取过程中的随机实验重复次数 $I = 500$. 此外, 算法 1 实现过程中的主要参数由表 1 给出. 接下来, 分别构建评判网络与执行网络激活函数为

$$\Phi(x, a) = [x^2(1), x(1)x(2), x(1)a, x^2(2), x(2)a, a^2]^{\mathrm{T}} \tag{33}$$

和

$$\Upsilon(x) = [x(1), x(2)]^{\mathrm{T}} \tag{34}$$

表 1 随机 Q-learning 算法的主要参数
Table 1 Main parameters of the stochastic Q-learning algorithm

算法参数	$\mathcal{Q}$	$\mathcal{R}$	$\rho_{\max}$	λ	ϵ
基准系统 I	$2I_2$	2.0	300	0.97	0.01
基准系统 II	$0.1I_4$	0.1	500	0.99	0.01

通过一个容许策略 $\mu(x) = -0.1x(1) - 3.5x(2)$, 初始化 Q 函数. 式 (9) 可通过不动点迭代技术[35]求解, 所得评判网络的初始权值为:

$$\vartheta_Q^{(0)} = [242.543\,8, -17.053\,5, -2.055\,6, 75.321\,3, 7.345\,9, 2.221\,7]^{\mathrm{T}} \tag{35}$$

相较于多步评估获得初始代价函数, 不动点迭代符合本文数据驱动思想. 执行随机系统 Q-learning 算法之后, 可得迭代 Q 函数权值曲线如图 1 所示. 当满足停止误差 $\epsilon = 0.01$ 后, 算法在 96 次迭代后收敛到近似最优值

$$\hat{\vartheta}_Q^* = \vartheta_Q^{(96)} = [115.289\,2, -2.111\,2, -0.672\,0, 28.917\,7, 2.686\,8, 2.080\,3]^{\mathrm{T}} \tag{36}$$

策略优化过程中, 执行网络的权值如图 2 所示. 综合图 1 和 2, 验证了本文所提出的随机系统 Q-learning 算法的收敛性.

分别应用初始策略 $\mu(x)$ 与算法 1 获得的近似最优策略 $\hat{\mu}^*(x)$ 后, 系统的状态曲线如图 3 所示, 其中, "In"代表初始策略下的系统响应, "Lim"代表近似最优策略下的系统响应, 细线代表系统运行过程中的期望, 包络线代表运行过程中的方差. 观察图 3

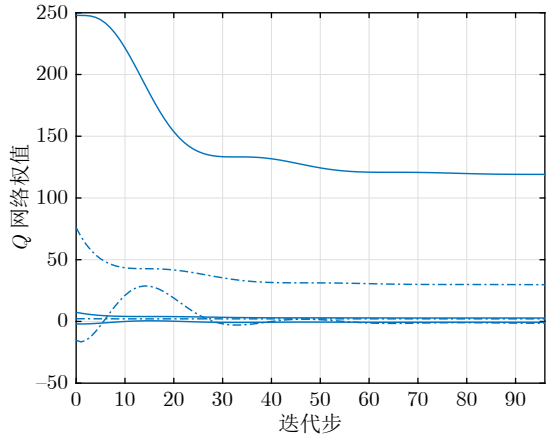


图 1 Q 网络权值曲线 (基准系统 I)
Fig. 1 Curves of Q network weights (Benchmark system I)

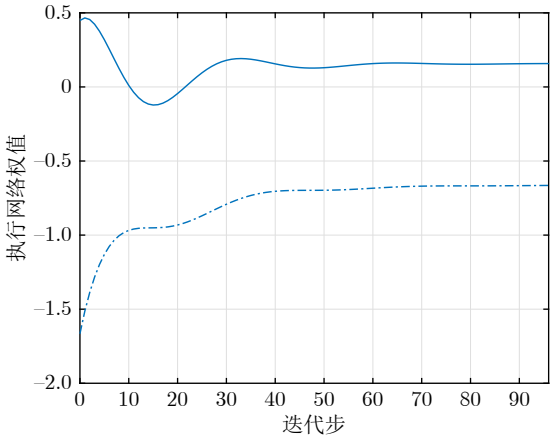
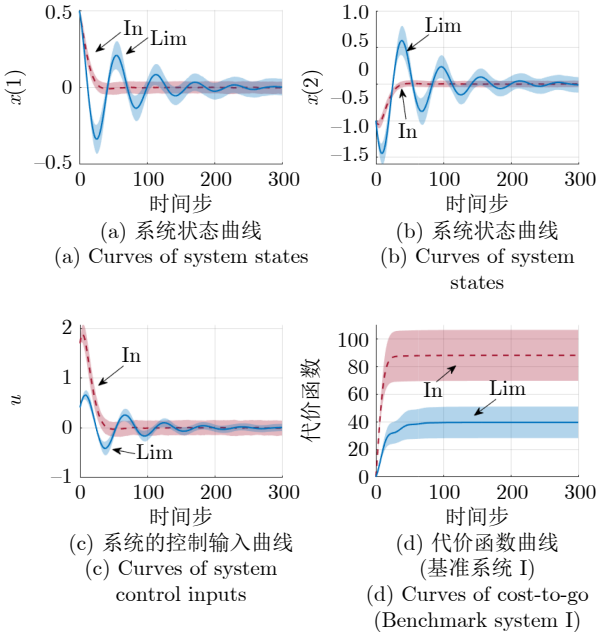


图 2 执行网络权值曲线 (基准系统 I)
Fig. 2 Curves of action network weights (Benchmark system I)

中子图 (a) 和 (b), 尽管系统在初始策略 $\mu(x)$ 下迅速将初始状态 $x_0 = [0.5, -0.5]^{\mathrm{T}}$ 镇定到平衡点附近, 但在综合考虑系统状态和控制成本的效用函数 $U(x, a) = x^{\mathrm{T}}\mathcal{Q}x + a^{\mathrm{T}}\mathcal{R}a$ 下是非最优的. 从图 3(d) 可以得到, 在折扣代价函数下, 随机系统 Q-learning 算法对初始策略进行优化, 并使得在近似最优策略 $\hat{\mu}^*(x)$ 下系统成本显著降低. 综上, 通过学习表现与系统运行表现, 充分验证了本文所提随机系统 Q-learning 算法的可行性与有效性.

4.2 基准系统 II

考虑在文献 [36] 中测试控制器性能的球台平衡系统, 该系统通过一个电机驱动平台控制球体的位置, 其结构如图 4 所示. 系统非线性动态通过如下方程给出:



(a) 系统状态曲线 (a) Curves of system states　(b) 系统状态曲线 (b) Curves of system states

(c) 系统的控制输入曲线 (c) Curves of system control inputs　(d) 代价函数曲线 (基准系统 I) (d) Curves of cost-to-go (Benchmark system I)

图 3　控制策略测试曲线 (基准系统 I)

Fig. 3　Curves of control policies for performance test (Benchmark system I)

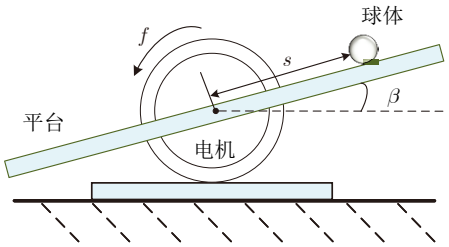


图 4　球台平衡系统示意图 (基准系统 II)

Fig. 4　Schematic diagram of the ball-and-beam system (Benchmark system II)

$$\begin{cases} \left(\dfrac{I_b}{\tau^2}+\varpi\right)\ddot{s}+\dfrac{I_b+\varpi\tau^2}{\tau}\ddot{\beta}-\varpi s\dot{\beta}^2=\varpi g\sin\beta \\ (\varpi s^2+I_b+I_\omega)\ddot{\beta}+(2\varpi s\dot{s}+f_cL^2)\dot{\beta}+ \\ \quad S_tL^2\beta+\dfrac{I_b+\varpi\tau^2}{\tau}\ddot{s}-\varpi gs\cos\beta= \\ \quad fL\cos\beta \end{cases} \tag{37}$$

其中, s 代表球体离电机中心的距离, β 代表平台与水平轴夹角, f 是电机的驱动力, 其他参数在表 2 给出. 选取系统的状态向量和控制输入分别为 $x=[s,\ \dot{s},\ \beta,\ \dot{\beta}]^{\mathrm{T}}$ 和 $u=f$. 由系统方程 (37), 可进一步得到球台平衡系统的状态空间方程. 对系统式 (37) 使用欧拉方法[37] 进行离散化, 并引入随机噪声, 离散状态空间方程可改写为

表 2　球台平衡系统的主要参数

Table 2　Main parameters of the ball-and-beam system

符号及取值	物理意义
$S_t=0.001$ N/m	驱动机械刚度
$L_\omega=0.5$ m	平台半径
$L=0.48$ m	电机作用半径
$f_c=1$ N$_\mathrm{s}$/m	驱动电机的机械摩擦系数
$I_\omega=0.140\ 25$ kg · m^2	平台惯性矩
$g=9.8$ m/s^2	重力加速度
$\varpi=0.016\ 2$ kg	球体质量
$\tau=0.02$ m	球体滚动半径
$I_b=4.32\times10^{-5}$ kg · m^2	球体转动惯量

$$x_{k+1}=\begin{bmatrix} x_k(1)+\Delta t x_k(2) \\ x_k(2)+1.717\Delta t\sin x_k(3) \\ x_k(3)+\Delta t x_k(4) \\ x_k(4)+\Delta t(-0.241x_k(4)+0.157x_k(1)\cos x_k(3)) \end{bmatrix}+ \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0.5\Delta t\cos x(3) \end{bmatrix}u_k+\begin{bmatrix} \omega_k \\ \omega_k \\ 0 \\ 0 \end{bmatrix}+\delta_k x_k \tag{38}$$

其中, 采样时间 $\Delta t=0.05$ s, $\omega_k\sim\mathrm{N}(0,\ 0.001^2)$ 和 $\delta_k\sim\mathrm{N}(0,\ 0.05^2)$ 是两个服从高斯分布的独立噪声.

基于算法 1, 首先从球台平衡系统中收集 7 776 组数据构成数据集 D_M, 其中 $\Omega_x=\{|x(1)|\le0.5,\ |x(2)|\le0.5,\ |x(3)|\le0.5,\ |x(4)|\le0.5\}$, $\Omega_\mu=\{|u|\le5\}$. 其余参数通过表 1 给出. 构建评判网络与执行网络激活函数为

$$\begin{aligned}\Phi(x,\ a)=[&x^2(1),\ x^2(2),\ x^2(3),\ x^2(4),\ x(1)x(2),\\ &x(1)x(3),\ x(1)x(4),\ x(2)x(3),\ x(2)x(4),\\ &x(3)x(4),\ x(1)a,\ x(2)a,\ x(3)a,\ x(4)a,\ a^2]^{\mathrm{T}}\end{aligned} \tag{39}$$

和

$$\Upsilon(x)=[x(1),\ x(2),\ x(3),\ x(4)]^{\mathrm{T}} \tag{40}$$

通过容许策略 $\mu(x)=-3x(1)-x(2)-15x(3)-2x(4)$ 初始化 Q 函数, 所得评判网络的初始权值为

$$\begin{aligned}\vartheta_Q^{(0)}=[&23.043\ 1,\ 43.006\ 6,\ 288.661\ 1,\ 49.702\ 6,\\ &20.195\ 6,\ 114.768\ 2,\ 23.362\ 3,\ 76.123\ 5,\\ &42.226\ 3,\ 121.274\ 1,\ 0.421\ 4,\ 0.922\ 5,\\ &2.109\ 1,\ 2.217\ 1,\ 0.125\ 9]^{\mathrm{T}}\end{aligned} \tag{41}$$

执行随机系统 Q-learning 算法, 获得迭代 Q 函数权值曲线如图 5 所示. 当满足停止误差 ϵ 后,

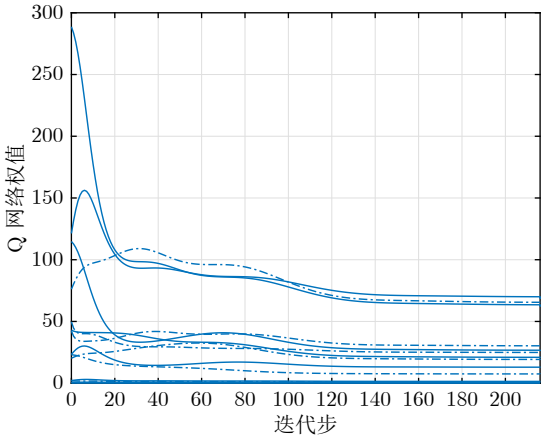


图 5 Q 网络权值曲线 (基准系统 II)

Fig. 5 Curves of Q network weights (Benchmark system II)

算法在 216 次迭代后收敛到近似最优值

$$\hat{\vartheta}_Q^* = \vartheta_Q^{(216)} = [8.740\,7,\ 25.191\,2,\ 74.640\,2,\ 27.226\,2,\ 23.174\,8,\ 31.818\,1,\ 14.725\,5,\ 78.280\,0,\ 34.818\,9,\ 79.311\,3,\ 0.314\,1,\ 0.738\,8,\ 1.637\,7,\ 1.195\,5,\ 0.113\,7]^{\mathrm{T}} \tag{42}$$

策略优化过程中, 执行网络的权值如图 6 所示.

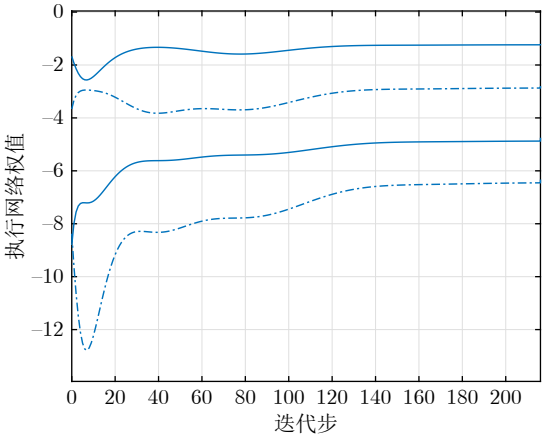


图 6 执行网络权值曲线 (基准系统 II)

Fig. 6 Curves of action network weights (Benchmark system II)

分别应用初始策略 $\mu(x)$ 与算法 1 中近似最优策略 $\hat{\mu}^*(x)$, 将随机球台平衡系统式 (38) 从初始点 $x_0 = [0.1,\ 0,\ 0.1,\ 0]^{\mathrm{T}}$ 镇定到平衡点邻域. 系统的状态和控制曲线分别如图 7 和 8 所示. 经过随机系统 Q-learning 算法的策略优化, 系统运行的状态响应与控制成本的期望均得到了显著优化. 如图 9 所示, 策略优化后的代价函数期望明显降低, 且优化后策略的代价函数方差更小. 基于此, 本文所提的随机

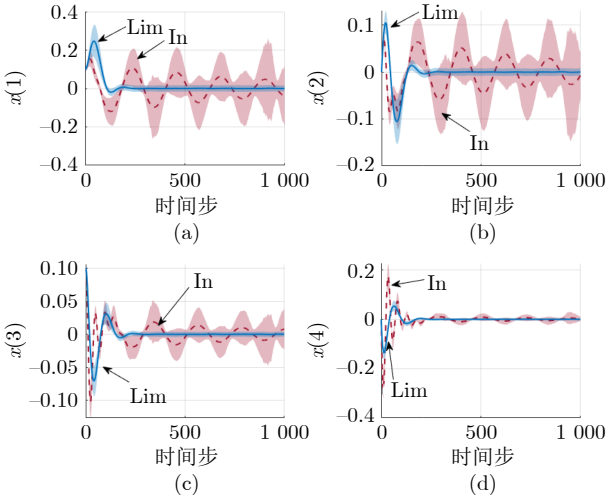


图 7 系统状态曲线 (基准系统 II)

Fig. 7 Curves of system states (Benchmark system II)

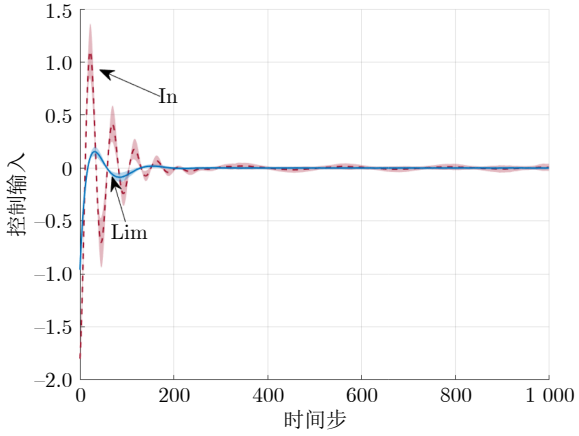


图 8 系统控制输入曲线 (基准系统 II)

Fig. 8 Curves of system control inputs (Benchmark system II)

系统 Q-learning 算法能够有效优化初始策略, 实现最小化代价函数的目标.

5 结束语

融合自适应评判技术, 本文建立了一种数据驱动的离散随机系统折扣最优调节方法. 理论证明了基于容许的初始策略, 迭代 Q 函数序列可以单调不增地收敛至最优. 此外, 我们给出了随机系统 Q-learning 算法的神经网络实施方案. 最后, 利用两个基准系统验证了本文所提出的随机系统策略优化算法. 事实上, 基于自适应评判技术, 数据驱动随机系统的工作仍有巨大的潜力. 未来可从如下几个方面深入探讨:

1) 本文算法的理论分析过程中, 不考虑迭代过程中神经网络近似误差. 然而, 在实际的分析中, 近

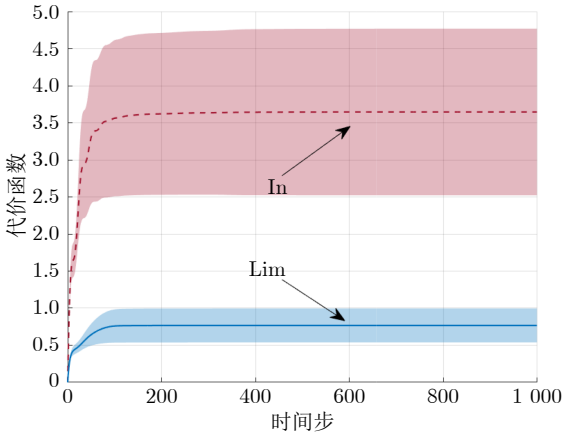


图 9　代价函数曲线 (基准系统 II)

Fig. 9　Curves of cost-to-go (Benchmark system II)

似误差可能会带来许多的影响, 包括系统的稳定性和算法的收敛分析. 值得一提的是, 由于随机系统需要在每一次迭代过程中重新拟合策略, 因此执行网络误差无法提供一种简化的分析思路, 如文献 [32] 所述的处理技巧. 这给误差分析带来了很大挑战.

2) 在算法的应用过程中, 可以利用收集的数据预训练系统模型. "模型"与"数据", "在线"与"离线"的混合学习, 可能会进一步提高算法效率以及控制性能.

3) 算法实现过程的主要难点是利用行为策略探索充分的数据, 数据是本文方法赖以执行的基础, 因此, 可考虑应用数据增强技术来提高自适应评判算法的综合性能, 例如, 典型的数字孪生、迁移学习等技术.

References

1　Liu D R, Xue S, Zhao B, Luo B, Wei Q L. Adaptive dynamic programming for control: A survey and recent advances. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2021, **51**(1): 142−160

2　Bellman R. Dynamic programming. *Science*, 1966, **153**(3731): 34−37

3　Wang F Y, Zhang H G, Liu D R. Adaptive dynamic programming: An introduction. *IEEE Computational Intelligence Magazine*, 2009, **4**(2): 39−47

4　Prokhorov D V, Wunsch D C. Adaptive critic designs. *IEEE Transactions on Neural Networks*, 1997, **8**(5): 997−1007

5　Zhao M M, Wang D, Qiao J F, Ha M M, Ren J. Advanced value iteration for discrete-time intelligent critic control: A survey. *Artificial Intelligence Review*, 2023, **56**(10): 12315−12346

6　Wang D, Gao N, Liu D R, Li J N, Lewis F L. Recent progress in reinforcement learning and adaptive dynamic programming for advanced control applications. *IEEE/CAA Journal of Automatica Sinica*, 2024, **11**(1): 18−36

7　Liu T, Tian B, Ai Y F, Li L, Cao D P, Wang F Y. Parallel reinforcement learning: A framework and case study. *IEEE/CAA Journal of Automatica Sinica*, 2018, **5**(4): 827−835

8　Miao Q H, Lv Y S, Huang M, Wang X, Wang F Y. Parallel learning: Overview and perspective for computational learning across Syn2Real and Sim2Real. *IEEE/CAA Journal of Automatica Sinica*, 2023, **10**(3): 603−631

9　Zhao M M, Wang D, Ha M M, Qiao J F. Evolving and incremental value iteration schemes for nonlinear discrete-time zero-sum games. *IEEE Transactions on Cybernetics*, 2023, **53**(7): 4487−4499

10　Wang Ding, Hu Ling-Zhi, Zhao Ming-Ming, Ha Ming-Ming, Qiao Jun-Fei. Event-triggered control design for optimal tracking of unknown nonlinear zero-sum games. *Acta Automatica Sinica*, 2023, **49**(1): 91−101
(王鼎, 胡凌治, 赵明明, 哈明鸣, 乔俊飞. 未知非线性零和博弈最优跟踪的事件触发控制设计. 自动化学报, 2023, **49**(1): 91−101)

11　Wang Ding. Event-based iterative neural control for a type of discrete dynamic plant. *Chinese Journal of Engineering*, 2022, **44**(3): 411−419
(王鼎. 一类离散动态系统基于事件的迭代神经控制. 工程科学学报, 2022, **44**(3): 411−419)

12　Wang D, Hu L Z, Zhao M M, Qiao J F. Dual event-triggered constrained control through adaptive critic for discrete-time zero-sum games. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2023, **53**(3): 1584−1595

13　Wang D, Li X, Zhao M M, Qiao J F. Adaptive critic control design with knowledge transfer for wastewater treatment applications. *IEEE Transactions on Industrial Informatics*, DOI: 10.1109/TⅡ.2023.3278875

14　Wang Ding, Zhao Hui-Ling, Li Xin. Adaptive critic control for wastewater treatment systems based on multi-objective particle swarm optimization. *Chinese Journal of Engineering*, 2024, **46**(5): 908−917
(王鼎, 赵慧玲, 李鑫. 基于多目标粒子群优化的污水处理系统自适应评判控制. 工程科学学报, 2024, **46**(5): 908−917)

15　Wu T Y, He S Z, Liu J P, Sun S Q, Liu K, Han Q L, et al. A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 2023, **10**(5): 1122−1136

16　Luo B, Liu D R, Wu H N, Wang D, Lewis F L. Policy gradient adaptive dynamic programming for data-based optimal control. *IEEE Transactions on Cybernetics*, 2017, **47**(10): 3341−3354

17　Luo B, Yang Y, Liu D R. Policy iteration Q-learning for data-based two-player zero-sum game of linear discrete-time systems. *IEEE Transactions on Cybernetics*, 2021, **51**(7): 3630−3640

18　Lin M D, Zhao B, Liu D R. Policy gradient adaptive critic designs for model-free optimal tracking control with experience replay. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2022, **52**(6): 3692−3703

19　Su S, Zhu Q Y, Liu J Q, Tang T, Wei Q L, Cao Y. A data-driven iterative learning approach for optimizing the train control strategy. *IEEE Transactions on Industrial Informatics*, 2023, **19**(7): 7885−7893

20　Wei Q L, Song R Z, Yan P F. Data-driven zero-sum neuro-optimal control for a class of continuous-time unknown nonlinear systems with disturbance using ADP. *IEEE Transactions on Neural Networks and Learning Systems*, 2016, **27**(2): 444−458

21　Liang M M, Wang D, Liu D R. Improved value iteration for neural-network-based stochastic optimal control design. *Neural Networks*, 2020, **124**: 280−295

22　Pang B, Jiang Z P. Reinforcement learning for adaptive optimal stationary control of linear stochastic systems. *IEEE Transactions on Automatic Control*, 2023, **68**(4): 2383−2390

23　Wei Q L, Zhou T M, Lu J W, Liu Y, Su S, Xiao J. Continuous-time stochastic policy iteration of adaptive dynamic programming. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2023, **53**(10): 6375−6387

24　Lee J, Haddad W M, Lanchares M. Finite time stability and optimal finite time stabilization for discrete-time stochastic dynam-

ical systems. *IEEE Transactions on Automatic Control*, 2023, **68**(7): 3978−3991

25 Liang M M, Wang D, Liu D R. Neuro-optimal control for discrete stochastic processes via a novel policy iteration algorithm. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2020, **50**(11): 3972−3985

26 Wang Ding, Zhao Ming-Ming, Ha Ming-Ming, Qiao Jun-Fei. Intelligent optimal tracking with application verifications via discounted generalized value iteration. *Acta Automatica Sinica*, 2022, **48**(1): 182−193
(王鼎, 赵明明, 哈明鸣, 乔俊飞. 基于折扣广义值迭代的智能最优跟踪及应用验证. 自动化学报, 2022, **48**(1): 182−193)

27 Wang D, Ren J, Ha M M, Qiao J F. System stability of learning-based linear optimal control with general discounted value iteration. *IEEE Transactions on Neural Networks and Learning Systems*, 2023, **34**(9): 6504−6514

28 Lincoln B, Rantzer A. Relaxing dynamic programming. *IEEE Transactions on Automatic Control*, 2006, **51**(8): 1249−1260

29 Ha M M, Wang D, Liu D R. Generalized value iteration for discounted optimal control with stability analysis. *Systems & Control Letters*, 2021, **147**: Article No. 104847

30 Ha M M, Wang D, Liu D R. Discounted iterative adaptive critic designs with novel stability analysis for tracking control. *IEEE/CAA Journal of Automatica Sinica*, 2022, **9**(7): 1262−1272

31 Yang X, Wei Q L. Adaptive critic designs for optimal event-driven control of a CSTR system. *IEEE Transactions on Industrial Informatics*, 2021, **17**(1): 484−493

32 Heydari A. Revisiting approximate dynamic programming and its convergence. *IEEE Transactions on Cybernetics*, 2014, **44**(12): 2733−2743

33 Ha M M, Wang D, Liu D R. Neural-network-based discounted optimal control via an integrated value iteration with accuracy guarantee. *Neural Networks*, 2021, **144**: 176−186

34 Wang D, Wang J Y, Zhao M M, Xin P, Qiao J F. Adaptive multi-step evaluation design with stability guarantee for discrete-time optimal learning control. *IEEE/CAA Journal of Automatica Sinica*, 2023, **10**(9): 1797−1809

35 Liu D R, Wei Q L. Policy iteration adaptive dynamic programming algorithm for discrete-time nonlinear systems. *IEEE Transactions on Neural Networks and Learning Systems*, 2014, **25**(3): 621−634

36 Zhong X N, Ni Z, He H B. Gr-GDHP: A new architecture for globalized dual heuristic dynamic programming. *IEEE Transactions on Cybernetics*, 2017, **47**(10): 3318−3330

37 Ha M M, Wang D, Liu D R. A novel value iteration scheme with adjustable convergence rate. *IEEE Transactions on Neural Networks and Learning Systems*, 2023, **34**(10): 7430−7442

王 鼎 北京工业大学信息学部教授. 2009 年获得东北大学硕士学位, 2012 年获得中国科学院自动化研究所博士学位. 主要研究方向为强化学习与智能控制. 本文通信作者.
E-mail: dingwang@bjut.edu.cn
(**WANG Ding** Professor at the Faculty of Information Technology, Beijing University of Technology. He received his master degree from Northeastern University in 2009 and Ph.D. degree from Institute of Automation, Chinese Academy of Sciences in 2012. His research interest covers reinforcement learning and intelligent control. Corresponding author of this paper.)

王将宇 北京工业大学信息学部博士研究生. 主要研究方向为强化学习和智能控制.
E-mail: wangjiangyu@emails.bjut.edu.cn
(**WANG Jiang-Yu** Ph.D. candidate at the Faculty of Information Technology, Beijing University of Technology. His research interest covers reinforcement learning and intelligent control.)

乔俊飞 北京工业大学信息学部教授. 主要研究方向为污水处理过程智能控制和神经网络结构设计与优化.
E-mail: adqiao@bjut.edu.cn
(**QIAO Jun-Fei** Professor at the Faculty of Information Technology, Beijing University of Technology. His research interest covers intelligent control of wastewater treatment processes, structure design and optimization of neural networks.)