数据驱动自适应评判控制研究进展

王 鼎 [1, 2, 3, 4] 赵明明 [1, 2, 3, 4] 刘德荣 [5] 乔俊飞 [1, 2, 3, 4] 宋世杰 [6]

摘 要 最优控制与人工智能的融合发展产生了一类以执行−评判设计为主要思想的自适应动态规划 (ADP) 方法. 通过集成动态规划理论、强化学习机制、神经网络技术、函数优化算法, ADP 在求解大规模复杂非线性系统的决策和调控问题上取得重要进展. 然而, 实际系统的未知参数和不确定扰动经常导致难以建立精确的数学模型, 对最优控制器的设计提出挑战. 近年来, 具有强大自学习和自适应能力的数据驱动 ADP 方法受到广泛关注, 它能够在不依赖动态模型的情况下, 仅利用系统的输入输出数据为复杂非线性系统设计出稳定、安全、可靠的最优控制器, 符合智能自动化的发展潮流. 通过对数据驱动 ADP 方法的算法实现、理论特性、相关应用等方面进行梳理, 着重介绍了最新的研究进展, 包括在线 Q 学习、值迭代 Q 学习、策略迭代 Q 学习、加速 Q 学习、迁移 Q 学习、跟踪 Q 学习、安全 Q 学习和博弈 Q 学习, 并涵盖数据学习范式、稳定性、收敛性以及最优性的分析. 此外, 为提高学习效率和控制性能, 设计了一些改进的评判机制和效用函数. 最后, 以污水处理过程为背景, 总结数据驱动 ADP 方法在实际工业系统中的应用效果和存在问题, 并展望一些未来的研究方向.

关键词 自适应评判控制, 自适应动态规划, 数据驱动设计, 在线 Q 学习, 迭代 Q 学习

引用格式 王鼎, 赵明明, 刘德荣, 乔俊飞, 宋世杰. 数据驱动自适应评判控制研究进展. 自动化学报, 2025, **51**(6): 1170−1190

DOI 10.16383/j.aas.c240706 **CSTR** 32138.14.j.aas.c240706

Research Advances on Data-driven Adaptive Critic Control

WANG Ding[1, 2, 3, 4] ZHAO Ming-Ming[1, 2, 3, 4] LIU De-Rong[5] QIAO Jun-Fei[1, 2, 3, 4] SONG Shi-Jie[6]

Abstract The fusion and development of optimal control and artificial intelligence yields adaptive dynamic programming (ADP) methods, which are primarily constructed based on the actor-critic design. By integrating dynamic programming theory, reinforcement learning mechanisms, neural network technologies, and function optimization algorithms, ADP has achieved significant progress in solving decision-making and control problems for large-scale complex nonlinear systems. However, the unknown parameters and uncertain disturbances of actual systems often make it difficult to establish accurate mathematical models, posing challenges to the design of optimal controllers. In recent years, data-driven ADP methods with strong self-learning and adaptive capabilities have received widespread attention. ADP methods can design stable, safe, and reliable optimal controllers for complex nonlinear systems using only the input-output data of the system without relying on dynamical models, aligning with the trend of intelligent automation. This paper comprehensively reviews the algorithm implementation, theoretical characteristics, and related applications of data-driven ADP methods, emphasizing the latest research progress, including online Q-learning, value-iteration-based Q-learning, policy-iteration-based Q-learning, accelerated Q-learning, transfer Q-learning, tracking Q-learning, safe Q-learning and game Q-learning. This paper also covers the analysis of data learning paradigms, stability, convergence, and optimality. Furthermore, in order to enhance learning efficiency and control performance, this paper designs some improved critic schemes and utility functions. Finally, with the background of wastewater treatment processes, this paper summarizes the application effects and existing issues of data-driven ADP approaches in practical industrial systems, and outlines several future research directions.

Key words Adaptive critic control, adaptive dynamic programming, data-driven design, online Q-learning, iterative Q-learning

Citation Wang Ding, Zhao Ming-Ming, Liu De-Rong, Qiao Jun-Fei, Song Shi-Jie. Research advances on data-driven adaptive critic control. *Acta Automatica Sinica*, 2025, **51**(6): 1170−1190

收稿日期 2024-10-31 录用日期 2025-01-17
Manuscript received October 31, 2024; accepted January 17, 2025
国家自然科学基金 (62222301, 62473012, 62021003), 国家科技重大专项 (2021ZD0112302) 资助
Supported by National Natural Science Foundation of China (62222301, 62473012, 62021003) and National Science and Technology Major Project (2021ZD0112302)
本文责任编委 陈谋
Recommended by Associate Editor CHEN Mou

1. 北京工业大学信息科学技术学院 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. 智慧环保北京实验室 北京 100124 4. 北京人工智能研究院 北京 100124 5. 南方科技大学自动化与智能制造学院 深圳 518055 6. 西南交通大学智慧城市与交通学院 成都 611756
1. School of Information Science and Technology, Beijing University of Technology, Beijing 100124 2. Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing 100124 3. Beijing Laboratory of Smart Environmental Protection, Beijing 100124 4. Beijing Institute of Artificial Intelligence, Beijing 100124 5. School of Automation and Intelligent Manufacturing, Southern University of Science and Technology, Shenzhen 518055 6. Institute of Smart City and Intelligent Transportation, Southwest Jiaotong University, Chengdu 611756

人工智能的蓬勃发展对科技革新和产业升级带来深远影响, 推动人类社会向着数字化、现代化、智能化的方向发展. 最优控制是指找到一个最优策略, 既能够镇定被控系统或受控对象, 同时使得设定的性能指标函数最小. 最优控制问题在现实生产生活中普遍存在, 涉及航空航天、系统工程、能源管理等许多领域[1]. 随着算法理论的发展和计算能力的提升, 人工智能技术在最优控制领域也得到广泛应用, 例如利用强化学习算法来优化控制策略, 赋予系统强大的自主学习和决策能力[2]. 强化学习是一种基于试错机制的智能化方法, 它强调智能体在与环境的交互过程中在线学习, 并研究其如何在环境中采取行动, 从而最大限度地增加累计奖励, 其中涉及最优化思想, 这与最优控制的核心理论不谋而合[3]. 因此, 强化学习被认为是解决复杂系统最优控制问题的重要技术之一.

在自适应控制领域, 利用强化学习思想与最优控制理论求解复杂系统优化控制问题的方法, 通常称为自适应动态规划 (Adaptive dynamic programming, ADP), 或者自适应评判设计[4]. 该方法首先基于动态规划中的最优性原理给出最优代价函数和最优控制策略形式, 然后利用强化学习中的执行–评判机制对控制策略进行反复的评价与更新, 使其逐渐逼近最优控制策略, 而神经网络通常作为算法实现过程中的函数近似工具[5]. 作为 ADP 方法中的三个重要组成部分, 代价函数形式一般与研究对象和实际问题相关, 评判学习机制通常对应求解方法和实现结构, 函数近似工具则关乎着目标策略的学习精度和算法复杂度. ADP 最大的优势在于集成多个领域的技术, 包括人工智能领域的强化学习算法和神经网络技术、控制领域的最优反馈与自适应调控技术、数学领域的迭代算法、计算智能领域的优化算法等. 一方面, 不同领域新技术的涌现可以促进 ADP 的进一步发展, 提升复杂系统的控制性能. 另一方面, ADP 也有助于改进已有的技术, 例如使用 ADP 代替传统的梯度下降算法来训练神经网络[6]. 总之, 在智能技术的日新月异发展下, 具有强大自学习和自适应能力的 ADP 更符合自动化系统设计的潮流, 在航天系统[7]、多智能体系统[8]、能源系统[9]、化工系统[10]、生物系统[11] 上具有广泛的应用前景.

纵观 ADP 方法的发展历程, 可以按照不同的标准进行分类. 考虑有无系统模型可分为基于模型的 ADP 方法和数据驱动的 ADP 方法, 考虑策略评估方式可分为值迭代 ADP 方法和策略迭代 ADP 方法, 考虑算法执行方式可分为在线学习 ADP 方法和离线学习 ADP 方法. 近年来的综述文献和学术专著, 对于基于模型的 ADP 方法, 在迭代算法的收敛性、控制策略的稳定性、方法实现的高效性等层面已进行详细总结[12−17]. 针对离散非线性系统的最优控制问题, Al-Tamimi 等[18] 在 2008 年提出一种初始条件为零的值迭代算法, 并证明了代价函数序列以单调非减的形式收敛到最优值. 这是第一次从数学上证明了零初始化值迭代算法的收敛性, 极大地推动迭代 ADP 算法的发展. 随后的十几年中, 学者们在初始化函数设计和值迭代更新过程方面做出大量的工作, 提出广义值迭代[19−22]、稳定值迭代[23]、局部值迭代[24]、演化值迭代[25]、集成值迭代[26]、多步值迭代[27−28]、λ 步值迭代[29−31]、并行值迭代[32]、平行值迭代[33]、可调节值迭代[34−35]、进化值迭代[36] 等一系列算法, 有效地减少了传统值迭代算法的计算量和保证了迭代过程中控制策略的稳定性. 与值迭代算法不同, 策略迭代算法要求一个容许策略进行初始化. 2014 年, Liu 等[37] 首次分析离散时间策略迭代算法的学习过程和相关特性, 证明了代价函数序列以单调非增的形式收敛到最优值, 这意味着迭代过程中的控制策略都是容许的. 在此基础上, 又衍生出性能更优的广义策略迭代[38]、局部策略迭代[39]、平衡策略迭代[40]、多步策略迭代[41]、λ-策略迭代[42]、加速策略迭代[43] 等算法. 值得一提的是, 这些基于模型的迭代 ADP 算法一般是离线实现的, 即通过离线迭代获得最优控制律后, 再将其作用于被控系统. 与之相对应的是基于模型的在线 ADP 算法, 它不再需要离线迭代, 而是直接采取在线自适应的方式获得最优控制律[44]. 目前, 基于模型的在线 ADP 算法在最优调节[45−46]、零和博弈[47]、跟踪控制[48−50]、事件触发控制[51−52]、鲁棒控制[53−54]、容错控制[55]、分散控制[56] 等方面已有比较丰富的研究成果.

上述 ADP 算法大多是基于模型的方法, 在过去的二十年中一直是该领域的研究热点. 然而, 在工程实际中, 未知的系统结构或参数经常给最优控制器的设计带来挑战. 为解决模型未知情况下的优化控制问题, 无模型或数据驱动的 ADP 方法受到广泛关注[57−61]. 实际系统在运行过程中会产生大量的过程数据, 这些离线和在线数据在一定程度上能够反映出系统的内在特性和运行规律. 在这些数据的赋能下, 数据驱动的 ADP 方法同样能够为复杂系统设计出稳定、安全、可靠的控制器. 数据驱动的 ADP 算法可分为两类, 即间接数据驱动的 ADP 算法和直接数据驱动的 ADP 算法. 间接数据驱动的 ADP 算法是指先利用数据建立近似的系统模型, 然后再进行控制器的设计和性能分析[62−64]. 相比之下,

直接数据驱动的 ADP 算法是指直接利用数据进行控制器的设计, 省去系统模型建立的过程[65−68]. 从本质上来说, 间接数据驱动的 ADP 算法的控制器设计利用近似系统模型, 所以相应的实现技术和理论特性与基于模型的 ADP 方法大致一样. 然而, 直接数据驱动的 ADP 算法则具有完全不同的实现结构, 理论特性也不能简单地移植. 因此, 本文将重点关注直接数据驱动的 ADP 算法. 事实上, 我们无意比较“基于模型的 ADP 算法”和“数据驱动的 ADP 算法”的性能优劣, 因为这两类算法在不同场景下具有各自的优势. 但对于系统模型难以建立且环境实时变化的情况, 数据驱动的 ADP 方法则展现出更广泛的应用前景和更大的发展潜力.

1989 年, Werbos 提出执行依赖启发式动态规划 (Action dependent heuristic dynamic programming, ADHDP) 结构[69]. 在 ADHDP 结构中, 评判网络的输入不仅有系统状态, 还有控制输入, 这意味着评判网络已经包含系统模型和效用函数的信息, 可直接通过最小化评判网络的输出得到控制律[70]. 因此, ADHDP 是一种数据驱动的 ADP 方法. 实际上, ADHDP 与 Watkins 博士论文中提出的 Q 学习算法在本质上为同一种结构, 因此 ADHDP 也称为 Q 学习[71]. Q 学习算法最初的研究主要集中在线性系统的最优控制理论和应用. 2007 年, Al-Tamimi 等[72] 提出一种基于值迭代的在线 Q 学习算法, 用于解决离散时间线性系统的零和博弈问题. 2014 年, Kiumarsi 等[73] 提出一种策略迭代 Q 学习方法, 在线实现了离散时间线性系统的数据驱动跟踪控制. 这两种 Q 学习算法为应用 ADP 方法解决线性零和博弈问题和跟踪控制问题设计了基本框架. 值得一提的是, 为确保对状态空间的充分探索, 通常需要加入探测噪声用于满足持续激励条件. 然而, 针对 Bellman 方程, 探测噪声可能会导致求解不准确. 为避免引入探测噪声带来的偏差, Jiang 等[74] 在 2012 年提出一种在线求解连续时间线性系统最优控制律的策略迭代算法, 核心是引入辅助变量对原系统进行重新描述, 这使得具有探测噪声的控制输入不会影响学习过程的收敛性和最优性. 在此基础上, 2017 年 Kiumarsi 等[75] 明确提出基于 off-policy 形式的强化学习算法, 有效实现了离散时间线性系统的在线 H_∞ 控制. 2022 年, Farjadnasab 等[76] 提出一种融合 off-policy 学习的无模型方法用于设计线性二次型调节器, 其过程主要是基于具有线性矩阵不等式约束的非迭代半定规划, 该方法的优势包括三个方面, 即提升采样效率, 对模型不确定性具有鲁棒性, 并且不需要初始稳定控制器. 2023 年, Lopez 等[77] 构建一种基于高效 off-policy 形式的 Q 学习算法, 与基于线性矩阵不等式解的控制设计方法相比, 其主要优点是降低了计算复杂度, 不需要初始稳定控制器, 并且对测量数据中的小干扰具有鲁棒性. 简单来说, 若产生数据的控制策略与目标策略不是同一个, 这样的学习形式称为 off-policy 学习. 反之, 若产生数据的控制策略与目标策略是同一个, 则称为 on-policy 学习. 目前, 普遍认为 off-policy 学习具有更好的探索能力. 近十年来, 许多面向线性系统的 Q 学习算法都以 off-policy 学习形式实现[78−83]. 随着线性系统数据学习范式的完善, 学者们将研究重点转向迭代算法的性能改进. 2023 年, Qasem 等[84] 提出混合迭代 Q 学习算法, 核心是利用值迭代提供初始容许策略, 然后使用策略迭代获得连续时间线性系统的最优控制策略. 2024 年, Jiang 等[85] 提出改进的 λ-策略迭代算法, 能够减少算法的迭代次数, 并给出严谨的收敛性分析. 2024 年, Zhao 等[86] 提出一种面向连续时间线性系统的单环策略迭代算法, 能够以更少的时间获得博弈代数 Riccati 方程的解. 如今, 面向线性系统的 Q 学习方法已经与控制理论和应用深度融合, 研究对象涉及博弈系统[87]、时滞系统[88−89]、随机系统[90]、执行器故障系统[91]、大规模系统[92]、网络化系统[93]、多智能体系统[94−96] 等, 应用层面包括机器人移动[97]、互联车辆控制[98]、微电网控制[99]、同步发电机控制[100] 等. 总的来说, 线性系统 Q 学习算法在收敛性、稳定性、鲁棒性、数据驱动范式、评判学习机制、实际应用等方面已具有相对成熟的工作.

在各行各业的控制应用中, 非线性系统普遍存在. 为实现最优控制, 一个传统的做法是将非线性系统简化或者近似为线性系统, 之后再利用已有的线性系统理论完成控制器的设计. 然而, 随着工业系统规模的扩大, 控制对象的复杂性不断增加, 并展现出强烈的非线性和未知性. 在这种情况下, 为模型未知的复杂非线性系统设计最优控制器, 对于确保系统正常运行并达到预定的性能指标具有重要意义. 早在 2001 年, Si 等[101] 就提出在线 Q 学习算法, 将当前时刻的代价函数和效用函数相加, 与上一时刻的代价函数作差, 通过最小化误差来实时更新评判网络权值. 然后, 根据最小化代价的原则调整执行网络的权值, 并将执行网络作为神经网络控制器, 实现对非线性系统的在线控制. 这种做法也常称为在线强化学习、在线 ADHDP 或者 direct HDP 方法. 2012 年, Liu 等[102] 以三层反向传播神经网络作为评判网络和执行网络的实现工具, 在假设输入层到隐藏层权值不变的前提下证明了在线 Q 学

习算法的一致最终有界性. 2015 年, Sokolov 等[103] 进一步给出输入层到隐藏层和隐藏层到输出层权值同时更新情况下的一致最终有界性. 这些工作都是基于目标策略产生的数据进行在线学习, 属于 on-policy 学习方法, 无法充分利用数据且需要较长的学习周期. 为提升学习效率, Malla 等[104] 将经验回放技术集成到传统的 ADHDP 方法中, 提高了算法学习的成功率和减少了训练时间. 从原理上讲, 能够利用过去的数据也属于 off-policy 学习方法. 总之, 上述在线 Q 学习算法重点研究控制器与系统实时交互场景下的系统稳定性和代价最优性.

近十年来, 关于非线性系统的迭代 Q 学习算法也涌现出大量的研究成果. 2014 年, Luo 等[105] 提出无模型的近似策略迭代算法, 通过提前收集连续时间系统的真实数据, 以迭代学习的方式获得近似最优控制律. 针对模型未知离散时间非线性系统的最优控制问题, Zhao 等[106] 在 2015 年提出一种需要初始容许控制的策略迭代 Q 学习算法, 并详细给出收敛性和稳定性分析, 重点指出每一个迭代策略都是容许的. 2024 年, Xu 等[107] 提出一种基于并行交叉熵优化方法的策略迭代 Q 学习算法, 通过求解二次规划问题获得初始容许的跟踪控制策略, 并证明了 Q 函数序列以单调非增的形式收敛到最优值. 2017 年, Wei 等[108] 提出确定的值迭代 Q 学习算法并严格证明了其收敛性. 2024 年, 王鼎等[109] 运用值迭代 Q 学习算法解决了离散时间随机系统的最优控制问题. 此后, Qiao 等[110] 建立基于加速值迭代的 Q 学习框架, 面对大量数据时能够以较小的计算代价获得最优 Q 函数. 这些方法都是利用提前采集的状态和控制数据对迭代 Q 函数进行更新, 从而获得目标迭代策略, 属于 off-policy 学习方法. 至今, 学者们对迭代 Q 学习算法的收敛性、稳定性、数据学习范式做了大量研究, 并给出不同应用场景下的算法实现结构. 然而, 目前尚没有文献对此进行归纳和凝练. 此外, 在开展实际的数据驱动控制过程中, 安全性、高效性、实用性也是需要考虑的重要特性. 安全性是指在不损伤系统的情况下, 仅利用历史数据获得最优控制策略. 高效性是指在面对大量数据时, 利用已有的知识或信息减少计算成本. 实用性是针对复杂工业系统, 降低控制器设计的复杂度. 针对未知环境下的一类最优控制问题, 本文将从经典的在线和迭代 Q 学习算法引入主题, 着重分析算法的理论特性和实际应用. 然后, 以实现高效学习和提升性能为目标, 提出一些新颖的 Q 学习算法. 最后, 展示一些典型应用场景并给出总结.

1 模型未知非线性系统的最优控制问题

考虑一类模型未知的离散时间非线性系统:

$$x_{k+1} = F(x_k,\ u_k),\ k = 0,\ 1,\ 2,\ \cdots \tag{1}$$

其中, $x_k = [x_k^{[1]},\ \cdots,\ x_k^{[j]}\cdots,\ x_k^{[n]}] \in \Omega_x \subset \mathbf{R}^n$ 是系统状态; $u_k = [u_k^{[1]},\ \cdots,\ u_k^{[j]}\cdots,\ u_k^{[m]}] \in \Omega_u \subset \mathbf{R}^m$ 是控制向量; $F(\cdot,\cdot)$ 是一个未知的非线性系统函数, 且 $F(0,\ 0) = 0$. 这里, $\mathbf{R}^n$ 表示由所有 n 维实向量组成的欧氏空间, Ω_x 和 Ω_u 是两个紧集. 假设非线性动态系统 (1) 可控. 注意, 式 (1) 中的系统通常称为输入非仿射系统, 与之对应的是输入仿射系统, 可表示为 $x_{k+1} = f(x_k) + g(x_k)u_k$, 其中 $f(\cdot)$ 和 $g(\cdot)$ 是未知的系统函数, 且 $f(0) = 0$.

针对离散时间系统的最优控制问题, 定义无限时域的代价函数为:

$$J(x_k) = \sum_{\ell=k}^{\infty} \gamma^{\ell-k} U(x_\ell,\ u_\ell) \tag{2}$$

其中, $0 < \gamma \leq 1$ 表示折扣因子; $U(x,\ u) = x^{\mathrm{T}}\mathcal{Q}x + u^{\mathrm{T}}Ru$ 表示与具体控制对象相关的效用函数, $\mathcal{Q}$ 和 R 是具有合适维数的对称正定有界矩阵. 根据 Bellman 最优性原理, 可通过求解如下离散 Hamilton-Jacobi-Bellman (HJB) 方程获得最优代价函数:

$$J^*(x_k) = \min_{u_k}\{U(x_k,\ u_k) + \gamma J^*(x_{k+1})\} \tag{3}$$

相应地, 最优控制策略可通过下式求解:

$$u^*(x_k) = \arg\min_{u_k}\{U(x_k,\ u_k) + \gamma J^*(x_{k+1})\} \tag{4}$$

注意到 $J^*(\cdot)$ 存在于式 (3) 的两边, 这导致 HJB 方程难以直接求解. 特别地, 当系统模型未知时, 求解 $J^*(\cdot)$ 将变得更加复杂. 为实现数据驱动的最优控制, 定义包含系统状态和控制输入的 Q 函数为:

$$\begin{aligned} Q(x_k,\ a) &= U(x_k,\ a) + \sum_{\ell=k+1}^{\infty} \gamma^{\ell-k} U(x_\ell,\ u_\ell) = \\ &U(x_k,\ a) + \gamma Q(x_{k+1},\ u_{k+1}) = \\ &U(x_k,\ a) + \gamma J(x_{k+1}) \end{aligned} \tag{5}$$

其中, $x_{k+1} = F(x_k,\ a)$; a 是一个任意的行为策略. 最优 Q 函数满足如下的最优方程:

$$\begin{aligned} Q^*(x_k,\ a) &= U(x_k,\ a) + \gamma \min_{u_{k+1}} Q^*(x_{k+1},\ u_{k+1}) = \\ &U(x_k,\ a) + \gamma J^*(x_{k+1}) \end{aligned} \tag{6}$$

最优控制律可表示为:

$$u^*(x) = \arg\min_{a} Q^*(x,\ a) \tag{7}$$

通过式 (7) 可以看出, 如果求出最优 Q 函数, 即可得到最优控制策略. 然而, 实际情况并非如此, 因为一般无法直接获得方程 (6) 的解析解. 为求解系统模型未知场景下的最优 Q 函数, 学者们提出在线 Q 学习算法和迭代 Q 学习算法, 这极大促进了数据驱动的 ADP 方法的发展. 值得注意的是, 无论在线学习或者迭代学习, Q 学习算法实现的前提是能够获得系统的输入输出数据. 令 $\{x_k,\ a,\ x_{k+1}\}$ 表示从系统 (1) 中采集的一组输入输出数据, 其中 $x_{k+1}=F(x_k,\ a)$. 令 $D_r=\{x_k^{(l)},\ a^{(l)},\ x_{k+1}^{(l)}\}_{l=1}^{L}$ 表示由多组输入输出数据构成的一个数据集, 其中 L 表示数据样本的个数. 特别地, 数据集可以是固定的, 也可以是变化的, 这对应着不同的 Q 学习算法和数据学习形式. 当前, 在线 Q 学习算法更注重实现结构的改进, 使其符合非线性系统设计实际应用. 相比之下, 迭代 Q 学习算法的理论仍处于持续发展阶段. 因此, 本文首先阐述在线 Q 学习算法的特性和应用, 然后对迭代 Q 学习算法进行较为全面的梳理.

2 非线性系统的在线 Q 学习算法

在线 Q 学习算法具有一些近义词, 如在线自适应评判、ADHDP、direct HDP 等, 但采用的实现结构都是一致的. 为了与迭代 Q 学习算法作出区分, 这里统称为在线 Q 学习算法, 其整体结构如图 1 所示, 其中包含用于近似控制律的执行网络, 用于近似 Q 函数的评判网络, 以及用于在线生成数据的真实系统.

定义具有折扣因子的 Q 函数为:

$$Q(x_k,\ u_k)=\sum_{\ell=k}^{\infty}\gamma^{\ell-k}U(x_{\ell+1},\ u_{\ell+1}) \tag{8}$$

可以看到, Q 函数已经包含状态 x_k 和控制输入 u_k 的信息, 这避免了对系统模型的依赖. 式 (8) 意味着 Q 函数满足:

$$0=\gamma Q_k-(Q_{k-1}-U_k) \tag{9}$$

其中, $Q_k=Q(x_k,\ u_k)$ 且 $U_k=U(x_k,\ u_k)$. 在线 Q 学习算法的核心是构建评判网络用于近似 Q 函数, 同时构建执行网络用于近似控制律 u_k. 评判网络和执行网络输出的近似值分别表示为:

$$\hat{Q}_k=W_{c2,\ k}^{\mathrm{T}}\Phi(W_{c1,\ k}^{\mathrm{T}}[x_k^{\mathrm{T}},\ \hat{u}_k^{\mathrm{T}}]^{\mathrm{T}}) \tag{10}$$

$$\hat{u}_k=W_{a2,\ k}^{\mathrm{T}}\Phi(W_{a1,\ k}^{\mathrm{T}}x_k) \tag{11}$$

其中, $W_{c1,\ k}$ 和 $W_{a1,\ k}$ 为输入层到隐藏层的权值, $W_{c2,\ k}$ 和 $W_{a2,\ k}$ 为隐藏层到输出层的权值, Φ 为激活函数. 根据式 (9), 定义评判网络的性能误差为:

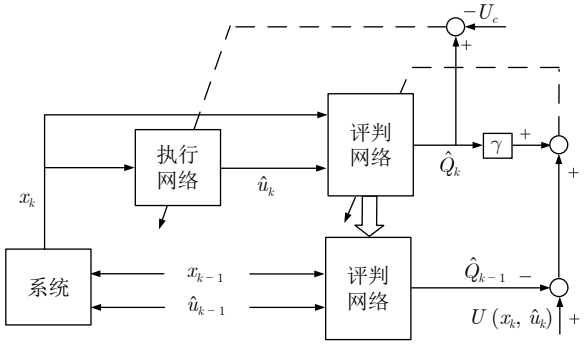

图 1 在线 Q 学习算法结构图

Fig. 1 The structure diagram of the online Q-learning algorithm

$$e_{c,\ k}=\gamma\hat{Q}_k+U(x_k,\ \hat{u}_k)-\hat{Q}_{k-1} \tag{12}$$

为使得目标函数 $E_{c,\ k}=0.5e_{c,\ k}^2$ 最小化, 一般采用梯度下降法来调整评判网络的权值. 定义执行网络的性能误差为 $e_{a,\ k}=\hat{Q}_k-U_c$, 其中 U_c 表示期望的代价成本, 通常设置为零. 为使得目标函数 $E_{a,\ k}=0.5e_{a,\ k}^{\mathrm{T}}e_{a,\ k}$ 最小化, 同样采用梯度下降法来更新执行网络的权值. 基于上述执行−评判结构和权值更新规则, 两个神经网络的权值能够在线实时更新. 随着时间的推移, 两个神经网络的权值最终收敛, 此时评判网络的输出逐渐逼近最优 Q 函数, 执行网络的输出逐渐逼近最优控制策略.

在线 Q 学习能够根据环境变化实时更新策略, 在算法特性和应用层面得到广泛研究. 设定 W_{c1} 和 W_{a1} 保持不变的情况下, 文献 [102] 证明了近似权值 (W_{c2} 和 W_{a2}) 与最优权值 (W_{c2}^* 和 W_{a2}^*) 之间的误差是一致最终有界的. 文献 [103] 对文献 [102] 中的工作进行扩展, 给出同时更新所有权值时的一致最终有界性条件. 为提升在线 Q 学习算法的数据利用效率, Ni 等[111] 将优先经验回放技术引入评判网络和执行网络的更新过程, 提高了多次独立运行时算法的成功率, 同时有效降低了算法的训练周期. 通过考虑未来的多步回报, Al-Dabooni 等[112] 提出一种具有长期预测参数的 HDP (λ) 方法, 在提高算法的学习效率和鲁棒性方面具有显著优势. 2022 年, 文献 [113] 将事件触发机制引入在线 Q 学习算法, 节省了通信资源且降低了随机波动的影响, 同时给出触发和不触发条件下系统的稳定性分析. 2021 年, Wei 等[114] 将在线 Q 学习技术应用到冰蓄冷空调系统中, 显著降低了系统的运行成本. 文献 [115] 利用在线 Q 学习算法实现了高炉煤气系统的粒度预测与动态调度, 实验结果表明该方法具有较高的精度. 对于一些多扰动非平稳的复杂动态系统, 如污

水处理过程, 若直接应用在线 Q 学习算法进行实时控制, 则可能导致灾难性的后果和造成经济损失. 因此, 一些学者尝试将先验知识引入在线 Q 学习算法, 保证系统稳定运行的前提下提升控制性能. 2024 年, 文献 [116] 提出融合知识迁移的自适应评判技术, 首先将离线评判学习得到的代价函数作为先验知识, 然后通过带有截断机制的衰减函数将其集成到在线 Q 学习设计中, 有效降低了试错成本且节约了计算资源. 随后, 文献 [117] 进一步将已有的经典控制器作为先验策略, 然后将其与在线 Q 学习控制策略进行集成, 实现了不同天气下污水处理过程中关键变量的精准控制. 事实上, 经验回放和先验知识的应用对提升非线性系统控制性能具有重要影响. 唯一不足的是, 由于在线数据不充分和神经网络近似不精确的原因, 在线 Q 学习算法的每一次运行结果可能呈现出随机性.

3　非线性系统的迭代 Q 学习算法

在大量数据的赋能下, 迭代 Q 学习算法的优势是能够保证每一次的运行结果一致. 根据迭代方式可将其分为值迭代 Q 学习和策略迭代 Q 学习算法. 本节主要讨论这两种算法的理论特性和数据学习形式.

3.1　值迭代 Q 学习算法

一类最基本的值迭代 Q 学习算法是利用固定的数据集 D 进行最优控制策略的学习. 由于系统 (1) 是确定的, 并且数据集内的状态 x_k 和动作 a 都是确定且不变的, 因此也被命名为确定的值迭代 Q 学习算法. 该算法由一个任意的半正定函数 $Q_0(x,\ a)$ 进行初始化, 随着迭代指标 $i=0,\ 1,\ 2,\ \cdots$ 的变化, 根据下式依次进行策略提升和 Q 函数更新:

$$u_i(x_{k+1})=\arg\min_a Q_i(x_{k+1},\ a) \tag{13}$$

$$Q_{i+1}(x_k,\ a)=\omega_i[U(x_k,\ a)+\gamma Q_i(x_{k+1},\ u_i(x_{k+1}))]+(1-\omega_i)Q_i(x_k,\ a) \tag{14}$$

其中, $0\le\omega_i\le 1$ 是与迭代指标相关的学习率. 需要注意的是, $u_i(x)$ 是期望得到的目标策略. 由于产生数据的行为策略 a 与目标策略 $u_i(x)$ 不是同一个策略, 这种学习方式称为 off-policy 学习. 确定的值迭代 Q 学习算法是一种离线学习方法, 具有初始化简单和容易实现的特点. 2017 年, Wei 等[108] 首次证明了该算法的收敛性, 并给出基于神经网络的算法实现结构. 2024 年, 文献 [110] 指出 ω_i 能够取大于 1 的有界常数值, 定义为 $\omega_i>1$, 由此提出一种可调节的值迭代 Q 学习算法, 实验结果表明可调节的 Q 函数序列比 $0\le\omega_i\le 1$ 情况下的 Q 函数序列收敛得更快. 由于数据驱动的迭代 Q 学习算法需要提前收集大量数据, 在算法学习过程中, 每一个迭代步都需要遍历所有数据, 这导致较大的计算压力. 相比之下, 具有加速优势的可调节值迭代 Q 学习算法能够通过减少迭代次数从而减少计算量, 属于改进的 Q 学习方法.

作为一种特例, 当 $\omega_i=1$ 和 $\gamma=1$ 时, Q 函数的更新过程简化为:

$$Q_{i+1}(x_k,\ a)=U(x_k,\ a)+Q_i\big(x_{k+1},\ u_i(x_{k+1})\big) \tag{15}$$

这也是目前应用最为广泛的一种形式. 算法的实现结构如图 2 所示, 其中执行网络用于逼近迭代控制策略, 近似值表示为 $\hat{u}_i(x)$, 评判网络用于逼近迭代 Q 函数, 近似值表示为 $\hat{Q}_{i+1}(x,\ a)$. 在迭代 Q 学习算法中, 神经网络和多项式是常用的近似工具. 随着控制对象和目标任务变得更加复杂, 算法对近似工具的精度要求也越来越高. 2018 年, 文献 [118] 证明了式 (15) 中 Q 函数序列的收敛特性, 指出不同的初始条件将导致迭代 Q 函数序列从不同的方向收敛到最优 Q 函数: 如果 $Q_0(x,\ a)\le Q_1(x,\ a)$, 则有 $Q_i(x,\ a)\le Q_{i+1}(x,\ a)$; 如果 $Q_0(x,\ a)\ge Q_1(x,\ a)$, 则有 $Q_i(x,\ a)\ge Q_{i+1}(x,\ a)$. 值得一提的是, 上述工作重点讨论 Q 函数序列的收敛性和单调性, 并没有研究迭代控制策略的稳定性. 2024 年, 文献 [119] 构建一个关于式 (15) 中 Q 函数序列的稳定性判别准则, 证明了如果两个连续的迭代 Q 函数满足:

$$Q_{i+1}(x_k,\ a)-Q_i(x_k,\ a)<\theta U(x_k,\ a) \tag{16}$$

其中, $0<\theta<1$, 则此时的控制策略 $u_i(x)$ 是容许的. 特别地, 如果 $Q_{i+1}(x,\ a)\le Q_i(x,\ a)$, 可得出单调非增的 Q 函数序列一定满足式 (16), 这意味着所有的迭代控制策略都是容许的. 值得一提的是, 由

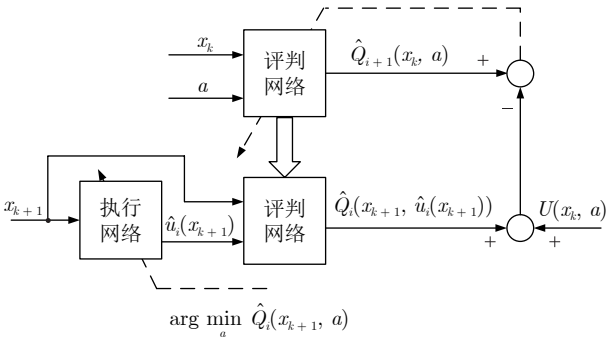

图 2　确定的值迭代 Q 学习算法结构图

Fig. 2　The structure diagram of the deterministic value iteration-based Q-learning algorithm

于满足式 (16) 的控制策略是容许的, 因此可作为策略迭代 Q 学习算法的初始容许策略. 除了实现系统 (1) 的最优调节之外, 值迭代 Q 学习框架也被推广用于解决非线性切换系统的控制问题[120−121].

与式 (15) 中的 off-policy 学习不同, on-policy 学习需要与系统进行交互, 它的 Q 函数更新过程为:

$$Q_{i+1}\big(x_k,\ u_i(x_k)\big) = U\big(x_k,\ u_i(x_k)\big) + Q_i\big(x_{k+1},\ u_{i-1}(x_{k+1})\big) \tag{17}$$

其中, $x_{k+1} = F(x_k,\ u_i(x_k))$. 由于产生数据的控制策略正是目标策略 $u_i(x_k)$, 这种学习方式称为 on-policy 学习. 随着迭代指标增大, 用于产生数据的控制策略 $u_i(x_k)$ 持续更新, 一旦进入到下一个迭代步, 当前迭代步的状态数据 x_{k+1} 就会被立即丢弃, 无法充分地利用历史数据. 针对离散时间仿射非线性系统, Li 等[122] 在 2019 年证明了 on-policy 学习框架下值迭代 Q 学习算法的收敛性, 同时指出 on-policy 学习的不足. 一方面, off-policy 学习中可以使用任意的行为策略来产生数据, 而 on-policy 学习中仅使用目标策略 $u_i(x_k)$, 因此 off-policy 学习具有更优的探索能力. 另一方面, 如果在 on-policy 学习中加入探测噪声来充分激励系统, 则会导致求解的最优控制策略产生偏差. 针对线性系统, 一种有效的改进方法是在 on-policy 学习过程中引入辅助变量, 使其转化成 off-policy 学习形式[75]. 基于这种思想, 文献 [122] 提出一种 off-policy 交错 Q 学习算法, 它的 Q 函数更新过程如下:

$$Q_{i+1}\big(x_k,\ u_i(x_k)\big) = U\big(x_k,\ u_i(x_k)\big) + Q_i\big(x^i_{k+1},\ u_{i-1}(x^i_{k+1})\big) \tag{18}$$

其中, $x^i_{k+1} = x_{k+1} - g(x_k)(a - u_i(x_k))$, 且 $x_{k+1} = F(x_k,\ a)$ 是提前获得的数据. 这种引入辅助变量的 off-policy 学习方法有效避免了探测噪声的影响. 2024 年, Song 等[123] 在考虑近似误差存在的情况下, 分析式 (18) 中 off-policy 学习框架下迭代 Q 函数的收敛性. 此外, 文献 [124] 将这种 off-policy 学习方法用于解决模型未知的离散时间非线性系统的跟踪控制问题, 通过对比实验阐明了该方法具有更好的鲁棒性. 然而, 该方法具有一定的局限性, 一是要求部分系统模型信息 $g(x_k)$, 二是仅适用于输入仿射系统. 在模型未知的情况下, 即使 $g(x_k)$ 可以用神经网络近似得到, 但对 $g(x_k)$ 的依赖使其无法推广到输入非仿射系统. 目前而言, 针对离散时间非线性系统, 式 (15) 中的 off-policy 学习应用更为广泛.

3.2 策略迭代 Q 学习算法

一类最基本的策略迭代 Q 学习算法也是利用固定的数据集 D 进行最优控制策略的学习, 称为确定的策略迭代 Q 学习. 令 $u_0(x)$ 表示一个初始容许控制策略. 对于迭代指标 $i = 0,\ 1,\ 2,\ \cdots$, 确定的策略迭代 Q 学习算法根据式 (19) 和 (20) 执行策略评估和策略提升:

$$Q_i(x_k,\ a) = U(x_k,\ a) + Q_i\big(x_{k+1},\ u_i(x_{k+1})\big) \tag{19}$$

$$u_{i+1}(x_{k+1}) = \arg\min_a Q_i(x_{k+1},\ a) \tag{20}$$

因为 a 与 $u_i(x)$ 不是同一个策略, 所以确定的策略迭代 Q 学习算法属于 off-policy 学习方法. 算法的实现结构如图 3 所示, 其中执行网络用于逼近迭代控制策略, 近似值表示为 $\hat{u}_i(x)$, 评判网络用于近似迭代 Q 函数, 近似值表示为 $\hat{Q}_i(x,\ a)$. 2015 年, 文献 [106, 125] 通过不同的数学方法证明了确定的策略迭代 Q 学习算法的收敛性, 重点指出迭代 Q 函数序列以单调非增的形式收敛到最优值, 即 $Q_{i+1}(x_k,\ a) \le Q_i(x_k,\ a)$, 且每一个迭代策略都是容许的. 在使用近似工具实现策略迭代 Q 学习算法时, 不可避免地会引入近似误差. 2017 年, 文献 [126] 给出迭代 Q 函数的误差界分析, 证明了在一个给定的有界条件下, 近似 Q 函数将收敛到最优 Q 函数的有限邻域. 2020 年, 文献 [127] 对策略迭代 Q 学习算法进行推广, 解决了非线性多智能体系统的最优控制问题.

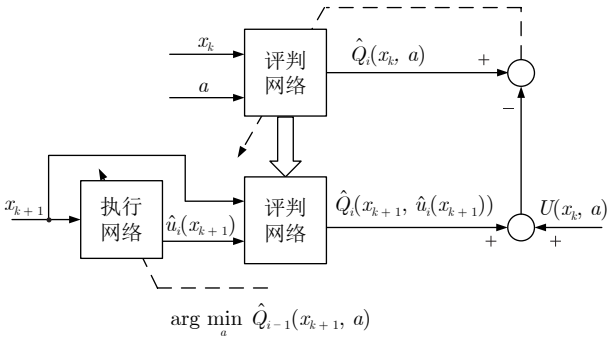

图 3 确定的策略迭代 Q 学习算法结构图

Fig. 3 The structure diagram of the deterministic policy iteration-based Q-learning algorithm

为了实现式 (20) 中的最小化, 文献 [128] 提出一种策略梯度 ADP 方法, 即通过 Q 函数关于控制策略求梯度来实现策略提升:

$$u_{i+1}(x) = u_i(x) - \alpha_c \frac{\partial Q_i(x,\ a)}{\partial a}\bigg|_{a=u_i(x)} \tag{21}$$

其中, $\alpha_c > 0$ 是学习率. 实际上, 由于每一个迭代策略都是容许的, 因此策略梯度 ADP 算法能够实现在线控制. 实现过程为将当前迭代步的控制策略作用于系统一个时间步, 然后将这个运行数据添加到

数据集中, 继续进行下一个迭代步的学习, 即迭代学习过程与控制过程随着时间同步进行. 尽管当前的运行数据是由目标策略产生, 具备 on-policy 学习的特征, 但由于数据集中包含以前的数据, 因此依然可视为 off-policy 学习[128]. 2020 年, 文献 [129] 提出一种确定的策略梯度 ADP 方法, 核心是引入深度 Q 学习中的经验回放和目标网络技术. 经验回放是指从数据集中随机选取小批量样本来更新神经网络权值, 避免陷入局部最优和发散. 目标网络是指固定执行网络和评判网络的权值, 使得神经网络的更新过程更快且更稳定. 2022 年, 文献 [130] 提出确定的孪生策略梯度 ADP 算法, 取两个 Q 网络之间的最小值来更新控制策略, 从而避免由函数逼近误差导致的过高评估问题. 2024 年, 文献 [131] 提出一种基于 OptNet 的策略梯度 ADP 方法, 首先使用非线性模型预测控制方法计算系统的输入输出轨迹, 然后通过 OptNet 获取策略迭代 Q 学习算法需要的初始容许控制策略.

与式 (19) 中的 off-policy 学习不同, 文献 [132] 提出一种基于 on-policy 学习的策略优化 ADP 方法, 其本质也是采用策略梯度技术来最小化 Q 函数, 它的策略评估过程如下所示:

$$Q_i\big(x_k,\ u_i(x_k)\big) = U(x_k,\ u_i(x_k)) + Q_i\big(x_{k+1},\ u_i(x_{k+1})\big) \tag{22}$$

借助 Polyak-Lojasiewicz 不等式, 文献 [132] 给出一种新的收敛性分析方法, 即迭代 Q 函数在有限迭代次数内能够收敛到一个给定阈值内的最优值. 由于式 (22) 是一种 on-policy 学习方式, 因此控制策略 $u_i(x_k)$ 需要与系统交互, 在每一个迭代步产生新的数据, 即 $x_{k+1} = F(x_k,\ u_i(x_k))$. 为充分利用数据, 文献 [133] 将 on-policy 学习过程中每个迭代步产生的数据添加到数据集中, 并丢弃一些旧数据, 从而以经验回放的形式实现 off-policy 学习.

至此, 我们总结了 Q 学习算法中常见的三种 off-policy 学习形式. 第一类是式 (15) 和 (19) 中所示的确定的迭代 Q 学习算法. 第二类是式 (18) 中所示的引入辅助变量法. 第三类是在 on-policy 学习过程中引入经验回放技术. 总之, 在数据驱动自适应评判控制领域, 我们认为能够利用历史数据的 Q 学习算法即为 off-policy 学习方法.

4 性能改进的 Q 学习算法

本节重点研究加速 Q 学习算法和迁移 Q 学习算法, 以提高学习效率和控制性能, 并设计一些新的评判学习机制.

4.1 加速 Q 学习算法

为降低迭代 Q 学习算法的计算代价和学习时间, 一种直观的做法是减少算法的迭代次数. 接下来, 我们分别从策略评估和策略提升两个角度出发, 讨论如何加快算法的学习速度.

策略评估的改进实际上是 Q 函数更新过程的改进. 在基于模型的 ADP 方法中, 一般认为策略迭代比值迭代具有更快的收敛速度, 这个结论同样适用于迭代 Q 学习算法. 因此, 为加快学习速度, 策略迭代 Q 学习算法受到广泛关注. 2018 年, 文献 [134] 提出一种自适应 Q 学习算法, 核心是通过引入一个自适应参数将值迭代 Q 学习和策略迭代 Q 学习算法进行结合:

$$Q_{i+1}(x_k,\ a) = U(x_k,\ a) + \alpha_i Q_{i+1}(x_{k+1},\ u_i(x_{k+1})) + (1-\alpha_i)Q_i\big(x_{k+1},\ u_i(x_{k+1})\big) \tag{23}$$

其中, $0 < \alpha_i < 1$. 当 $\alpha_i \equiv 0$ 时, 自适应 Q 学习算法变化为式 (15) 中的值迭代 Q 学习算法, 不要求初始容许控制策略. 当 $\alpha_i \equiv 1$ 时, 自适应 Q 学习算法变化为式 (19) 中的策略迭代 Q 学习算法, 具有较快的收敛速度但需要一个初始容许控制策略. 因此, 自适应 Q 学习算法通常以一个较小的 α_i 开始执行, 从而避免初始容许控制策略, 然后逐渐增大 α_i 以加快算法的收敛速度. 另外一种常见的做法是先采用值迭代 Q 学习算法获得一个初始容许控制策略, 然后开始执行策略迭代 Q 学习算法, 这种方法被称为混合迭代 Q 学习算法[135]. 然而, 对于非线性系统, 策略迭代 Q 学习算法本身就会引入较大的计算压力, 这是因为在每一个迭代步都包含一个完整的内层迭代, 需要连续的 Q 函数近似. 换句话说, 策略迭代 Q 学习算法只是减少迭代次数, 但不意味着减少计算时间和计算代价. 因此, 直接对值迭代 Q 学习算法进行加速也是一种有效的手段. 通过将历史迭代信息集成到当前迭代步, 提出一种可调节的值迭代 Q 学习算法:

$$Q_{i+1}(x_k,\ a) = U(x_k,\ a) + Q_i\big(x_{k+1},\ u_i(x_{k+1})\big) + \theta\Big[U(x_k,\ a) + Q_i\big(x_{k+1},\ u_i(x_{k+1})\big) - Q_\varkappa(x_k,\ a)\Big] \tag{24}$$

其中, $\theta \geq 0$ 是一个有上界的加速因子; $\varkappa$ 是一个正整数, 取值范围通常为 $\{i,\ i-1\}$. 当迭代 Q 函数序列为单调非减时, 可调节的值迭代 Q 学习算法相当于在式 (15) 的基础上增加一个正数项 $\theta\big[U(x_k,\ a) +$

$Q_i(x_{k+1},\ u_i(x_{k+1})) - Q_{\varkappa}(x_k,\ a)\big]$, 因此 Q 函数能够更快地接近最优值, 从而实现加速学习. 值得一提的是, $\varkappa = i$ 是较为常见的选择, 并且能够从理论上保证算法的收敛性[110]. 相比之下, $\varkappa$ 的取值为 $i-1$ 时能够获得更快的学习速度, 但收敛性尚缺乏保证. 为兼顾快速性和收敛性, 可以采用切换的方法, 首先使用可调节的值迭代 Q 学习算法进行加速, 然后切换为传统的值迭代 Q 学习算法用于保证收敛性.

策略提升的改进实际上是控制策略求解过程的改进. 从本质上看, 寻找能够最小化 Q 函数的控制策略是一个优化问题. 因此, 可以将优化领域的一些经典和前沿算法用于寻找每个迭代步的控制策略. 梯度下降算法由于简单易实现的特点, 已广泛用于求解最小化 Q 函数的控制策略:

$$u_{i,\,q}(x) = u_{i,\,q-1}(x) - \alpha_c \frac{\partial Q_i(x,\ a)}{\partial a}\bigg|_{a=u_{i,\,q-1}(x)} \tag{25}$$

其中, $u_{i,\,0}(x) = u_{i-1}(x)$; $\alpha_c > 0$ 是学习率; q 是梯度下降的次数. 在经历合适的更新次数后, 将最终得到的 $u_{i,\,q}(x)$ 记为提升后的迭代策略 $u_i(x)$. 然而, 对于一些具有实际应用背景的复杂 Q 函数, 式 (25) 中的传统梯度下降法需要较多的次数才能找到理想的控制策略. 为改善这个过程, 可以利用一些改进的梯度算法来实现策略提升, 例如牛顿法、拟牛顿法、Nesterov momentum、Adaptive gradient (AdaGrad)、Adadelta、Root mean square prop (RMSprop)、Adaptive moment estimation (Adam) 等. 以常见的 Nesterov momentum 法为例, 这里给出最小化 Q 函数时的控制策略求解过程:

$$\begin{cases} B_q = u_{i,\,q}(x) + \beta A_q \\ A_{q+1} = \beta A_q - \alpha_c \dfrac{\partial Q_i(x,\ B_q)}{\partial B_q} \\ u_{i,\,q+1}(x) = u_{i,\,q}(x) + A_{q+1} \end{cases} \tag{26}$$

其中, A_q 和 B_q 是与迭代次数 q 相关的中间变量, $A_0 = 0$; $\beta > 0$ 是动量参数. 值得一提的是, 上述优化方法至今仍是强化学习和深度学习领域的研究热点, 前沿算法层出不穷, 如 Adams[136]、AngleAdam[137]、MonAdam[138] 等方法, 都是经典 Adam 算法的改进, 将其与 ADP 相结合符合智能优化控制的发展潮流. 除了梯度算法, 群智能算法也是解决复杂优化问题的有力工具. 作为群智能算法的一个重要分支, 粒子群优化算法具有简单易行、收敛速度快、全局搜索能力强等优点, 被广泛应用于函数优化和神经网络训练等领域. 2023 年, 文献 [139] 尝试用粒子群优化算法来更新评判网络权值, 实现了连续时间非线性互联系统的事件触发局部控制. 2024 年, 针对离散时间系统, 文献 [36] 采用粒子群优化算法更新执行网络权值, 解决了系统状态对控制输入不可导情况下的最优控制问题. 实际上, 文献 [36, 139] 并没有直接利用粒子群优化算法来寻找控制律, 而是用它来更新神经网络, 但依然为推动群智能算法和 ADP 结合提供了重要技术指导. 在数据驱动自适应评判控制领域, 对于控制策略准确性要求更高的场景, 先进的群智能算法无疑具有更大的应用前景.

4.2 迁移 Q 学习算法

为提升新任务的学习性能, 将迁移学习与 ADP 相结合已成为自适应评判控制领域的研究热点. 2024 年, 文献 [116] 提出一种面向污水处理过程的在线迁移启发式动态规划算法, 利用收集的历史数据建立模型网络, 并将基于模型的评判学习得到的代价函数作为先验知识. 随后, 文献 [140] 提出一种针对不确定非线性系统的知识迁移自适应评判控制方法, 不仅考虑先验代价函数, 同时在学习过程中增加扰动补偿控制. 然而, 这些在线迁移 ADP 方法的先验知识都需要真实的或近似的系统模型. 文献 [141−142] 提出两种与迭代 Q 学习相关的迁移强化学习方法, 使设计的新控制器能够利用从先前的学习任务和数据中提取出的先验知识. 文献 [141] 采用值迭代 Q 学习算法, 其 Q 函数更新过程如下所示:

$$Q_{i+1}(x_k,\ a) = U'(x_k,\ a) + \gamma Q_i\big(x_{k+1},\ u_i(x_{k+1})\big) \tag{27}$$

其中, $U'(x_k,\ a) = U(x_k,\ a) + U_f(x_k,\ a)$, $U_f(x_k,\ a) = Q'(x_k,\ a) - \gamma Q'(x_{k+1},\ u_{k+1})$; Q' 表示从源任务中获得的先验知识. 文中也给出迁移强化学习框架下值迭代 Q 学习算法的收敛性和最优性分析, 并通过倒立摆小车杆平衡控制和人机交互的机器人假肢控制验证了算法的有效性. 文献 [142] 提出灵活策略迭代 Q 学习算法, 其 Q 函数更新过程为:

$$\begin{aligned} Q_i(x_k,\ a) &= U(x_k,\ a) + s_i\mathcal{V}(x_k) + \\ &\sum_{\ell=1}^{\infty}\big[U\big(x_{k+\ell},\ u_i(x_{k+\ell})\big) + s_i\mathcal{V}(x_{k+\ell})\big] = \\ &U(x_k,\ a) + s_i\mathcal{V}(x_k) + Q_i\big(x_{k+1},\ u_i(x_{k+1})\big) \end{aligned} \tag{28}$$

其中, s_i 表示补充系数, 满足 $0 \le s_{i+1} < s_i < 1$ 和 $\lim_{i\to\infty} s_i = 0$; $\mathcal{V}(x)$ 表示从之前实验中获得的补充值, 且 $\mathcal{V}(x) = \min_a Q_f(x_k,\ a)$, $Q_f(x_k,\ a)$ 表示没有补充值情况下的策略迭代算法的收敛值. 文中给出灵活策略迭代算法的收敛性和最优性分析, 同时引

入优先经验回放技术来提升策略评估过程. 通过在机器人膝关节上开展实验, 结果表明迁移强化学习方法在解决具有高维控制输入的现实问题时具有巨大潜力.

值得一提的是, 上述方法只迁移或利用已有的代价函数, 并没有考虑与控制策略相关的先验知识. 从本质上来说, 补充控制也属于迁移学习的范畴, 即将已有的经典控制器作为先验策略, 然后将 ADP 的控制策略作为补充策略, 这样能够提升先验策略的自适应学习能力并减少自适应评判学习所需的训练成本. 2016 年, 文献 [143] 设计一种附加控制器, 以双馈风机风电场的附加无功控制为例进行仿真测试, 验证了附加学习控制方法的在线优化能力和对不确定性的适应能力. 2024 年, 文献 [117] 提出一种具有经验回放的 ADHDP 补充控制器, 其中 PID 作为先验策略保证污水处理过程的稳定运行, ADHDP 控制器作为补充策略进一步提升控制性能. 值得一提的是, 文献 [117, 143] 中的方法都未将先验策略纳入 Q 函数更新过程中, 而是将在线学习得到的补充策略与先验策略进行简单的相加操作. 为考虑先验策略的影响, 本文提出一种集成 Q 学习算法, 其 Q 函数形式为:

$$Q(x_k,\ \lambda\mathcal{L}_k+\mathcal{B}_k)=\sum_{\ell=k}^{\infty}\gamma^{\ell-k}U(x_{\ell+1},\ \lambda\mathcal{L}_{\ell+1}+\mathcal{B}_{\ell+1}) \tag{29}$$

其中, $\mathcal{L}_k$ 为先验策略; $\mathcal{B}_k$ 为探索策略; λ 为比例系数, 决定了先验策略在整个控制律中的比重. 先验策略已在实际运行中得到检验, 在线学习过程中不再被更新. 探索策略是一种强化学习策略, 通过与环境交互收集的新经验进行在线训练. 对于一些复杂的工业系统, 这种集成学习方式能够使得探索策略在一个相对稳定的场景下进行学习, 并利用其在线自趋优和自适应能力持续提高控制性能[144]. 值得注意的是, 集成学习既适用于在线 Q 学习算法, 也适用于迭代 Q 学习算法. 总而言之, 将先验策略集成到评判学习过程中, 能够降低不稳定策略与系统交互带来的风险, 同时降低了控制器的设计难度.

5 效用改进的 Q 学习算法

除了在最优调节问题中取得重要进展之外, 在线 Q 学习和迭代 Q 学习算法也广泛应用于解决非线性系统的跟踪问题、约束问题、博弈问题等. 面向不同的被控对象, 通常需要设计新的效用函数和代价函数, 这也是本文的研究重点. 值得一提的是, 加速和迁移等性能改进的 Q 学习机制也能够推广到跟踪、约束、博弈等场景, 实现高效学习.

5.1 跟踪 Q 学习算法

最优跟踪控制的目标是找到一个反馈控制策略 u_k, 确保式 (1) 中的系统状态跟踪上有界的参考轨迹, 同时最小化与跟踪误差相关的性能指标函数. 这里, 定义参考轨迹系统为:

$$r_{k+1}=\mathcal{R}(r_k) \tag{30}$$

其中, $\mathcal{R}(\cdot)$ 是一个连续的函数. 定义跟踪误差为 $e_k=x_k-r_k$, 由此可得跟踪误差系统为:

$$e_{k+1}=F(e_k+r_k,\ u_k)-\mathcal{R}(r_k) \tag{31}$$

为实现系统模型未知情况下的最优跟踪控制, 前提是能够获得原系统的数据和参考系统的数据, 并建立一个包含跟踪误差和参考轨迹的数据集 $D_t=\{e_k^{(l)},\ r_k^{(l)},\ a^{(l)},\ e_{k+1}^{(l)},\ r_{k+1}^{(l)}\}_{l=1}^{L}$. 接下来, 我们重点介绍 ADP 跟踪控制方法中常用的四类效用函数形式:

$$\begin{cases}U_1(e_k,\ u(e_k))=e_k^{\mathrm{T}}\mathcal{Q}e_k+u^{\mathrm{T}}(e_k)Ru(e_k)\\ U_2(e_k,\ r_k,\ a)=e_k^{\mathrm{T}}\mathcal{Q}e_k+a^{\mathrm{T}}Ra\\ U_3(e_k,\ r_k,\ a)=e_{k+1}^{\mathrm{T}}\mathcal{Q}e_{k+1}\\ U_4(e_k,\ r_k,\ a)=e_k^{\mathrm{T}}\mathcal{Q}e_k+\Delta e_{k+1}^{\mathrm{T}}R\Delta e_{k+1}\end{cases} \tag{32}$$

其中, 第一类效用函数中的 $u(e_k)$ 是反馈控制, 它与前馈控制 $u(r_k)$ 相加组成控制策略 $u(x_k)$, 即 $u(x_k)=u(e_k)+u(r_k)$. 前馈控制也称为稳态控制或者参考控制, 用于实现完美跟踪 $x_{k+1}=r_{k+1}$, 因此可根据 $r_{k+1}=F(r_k,\ u(r_k))$ 求解获得. 此外, 第四类效用函数中的误差差分 $\Delta e_{k+1}=e_{k+1}-e_k$.

对于模型已知的仿射非线性系统, 最常用的 ADP 跟踪方法是采用第一类效用函数, 通过 $u(r_k)=g^{+}(r_k)(r_{k+1}-f(r_k))$ 求得前馈控制, 其中 $g^{+}(r_k)$ 是 $g(r_k)$ 的广义逆矩阵[145]. 事实上, 当系统模型未知时, 仍然可以通过神经网络建模来求解前馈控制[146]. 然而, 如果前馈控制不存在, 这些基于真实模型或近似模型的方法则会失效.

为避免求解前馈控制, 数据驱动跟踪 Q 学习算法受到广泛关注. 2016 年, 文献 [147] 提出策略迭代跟踪 Q 学习算法, 利用第二类效用函数解决了离散时间非线性系统的跟踪控制问题, 其中 Q 函数的更新过程如下所示:

$$\begin{aligned}Q_i(e_k,\ r_k,\ a)=U_2(e_k,\ r_k,\ a)+\\ \gamma Q_i\big(e_{k+1},\ r_{k+1},\ u_i(e_{k+1},\ r_{k+1})\big)\end{aligned} \tag{33}$$

随着迭代指标增大, 跟踪 Q 函数序列以单调非增的形式收敛到最优 Q 函数, 即 $Q_{i+1}(e_k,\ r_k,\ a)\le$

$Q_i(e_k,\ r_k,\ a)$. 由于控制策略 a 是任意的, 因此该算法属于 off-policy 学习, 算法所需的数据都可以从数据集 D_t 中获得. 随后, 文献 [133] 在式 (33) 的基础上引入策略梯度和经验回放技术, 进一步提升了数据利用效率且降低了策略求解的复杂度. 从式 (33) 可以看出, 迭代控制策略 $u_i(e,\ r)$ 与跟踪误差 e 和参考轨迹 r 相关. 因此, 在与时间相关的控制过程中, 控制策略的形式为 $u_k = u(e_k,\ r_k)$, 这意味着如果 $r_\infty \neq 0$, 则 $u_\infty \neq 0$, 于是在没有折扣因子的情况下会导致代价函数无界. 也就是说, 只有当参考轨迹是渐近稳定的, 即 $r_\infty = 0$, 折扣因子才能够取 1. 此外, 文献 [148] 指出使用 $U_2(e_k,\ r_k,\ a)$ 的跟踪方法无法完全消除跟踪误差, 这是因为控制输入 u_k 的最小化无法保证跟踪误差 e_k 的最小化.

为不限制参考轨迹形式且消除最终跟踪误差, 文献 [148] 设计第三类效用函数 $U_3(e_k,\ r_k,\ a)$, 直接省去控制输入的二次型. 在此基础上, 文献 [149] 利用基于模型的值迭代算法获取最优跟踪策略, 解决了离散时间非线性系统的跟踪控制问题. 2024 年, 文献 [150] 将 $U_3(e_k,\ r_k,\ a)$ 推广到值迭代跟踪 Q 学习算法, 其 Q 函数更新过程如下:

$$Q_{i+1}(e_k,\ r_k,\ a) = U_3(e_k,\ r_k,\ a) + Q_i\big(e_{k+1},\ r_{k+1},\ u_i(e_{k+1},\ r_{k+1})\big) \tag{34}$$

应该看到, 效用函数 $U_3(e_k,\ r_k,\ a)$ 只与下一时刻的跟踪误差有关, 因此 $e_\infty = 0$ 能够保证代价函数的有界性. 文献 [150] 分析了算法的收敛性和最优性, 并讨论了近似器误差存在情况下的策略稳定性. 为减少计算代价, 文献 [151] 在式 (34) 的基础上引入加速学习机制, 构建了加速的跟踪 Q 函数形式:

$$\begin{aligned} Q_{i+1}\ (e_k,\ r_k,\ a) = \zeta\Big[&U_3(e_k,\ r_k,\ a) + \\ &Q_i\big(e_{k+1},\ r_{k+1},\ u_i(e_{k+1},\ r_{k+1})\big)\Big] + \\ &(1-\zeta)Q_i(e_k,\ r_k,\ a) \end{aligned} \tag{35}$$

其中, $\zeta > 1$ 是一个有上界的加速因子. 值得一提的是, 由于效用函数 $U_3(e_k,\ r_k,\ a)$ 中不包含控制输入, 常规的构造非二次型函数克服控制约束的方法变得不可用. 基于系统转换技术, 文献 [152] 提出一种广义的性能指标函数, 核心是对任意的行为策略 a 施加约束得到新的控制输入 $u = \Psi(a)$, 其中 $\Psi(\cdot)$ 是一个约束函数, 然后通过 $x_{k+1} = F(x_k,\ \Psi(a))$ 生成数据. 这样学习过程中仍然采用不受约束的行为策略 a, 仅仅在控制过程中使用 $u = \Psi(a)$ 来实现不对称约束情况下的跟踪控制.

2023 年, 针对连续时间非线性系统的跟踪控制问题, 文献 [153] 提出新的代价函数形式 $J(e(0)) = \int_0^\infty U(e(\tau),\ \dot{e}(\tau))\mathrm{d}\tau$, 其中 $U(e,\ \dot{e}) = e^{\mathrm{T}}\mathcal{Q}e + \dot{e}^{\mathrm{T}}R\dot{e}$. 除了能够完全消除跟踪误差和不需要折扣因子的优势外, 该效用函数形式不要求原系统满足 $f(0) = 0$ 的假设[154]. 鉴于此, 本文将其推广到离散时间的跟踪 Q 学习算法, 给出第四类效用函数形式 $U_4(e_k,\ r_k,\ a)$. 值得注意的是, 基于 $U_4(e_k,\ r_k,\ a)$ 的跟踪 Q 学习算法的实现过程和理论特性与已有的跟踪 Q 学习算法保持一致, 数据也可从集合 D_t 中获得.

5.2 安全 Q 学习算法

对于一些与安全相关的实际系统, 如自动驾驶车辆和机器人移动, 安全性是至关重要的. 在设计最优控制器时, 不安全的探索行为可能会导致难以恢复的灾难性后果. 因此, 设计基于 ADP 的安全最优控制器, 确保最优性并满足安全约束, 是一项富有挑战性且具有实际意义的工作. 在自适应评判控制领域, 常见的安全问题主要包含输入约束和状态约束. 对于具有输入约束的非线性系统最优控制, 目前已有大量的研究成果. 常规的做法是在代价函数中设计一个关于控制输入的积分项, 使得求解出的控制输入满足约束范围[155]. 此外, 还有一种做法是通过系统转换技术, 将受输入约束的最优控制问题转换为无约束的最优控制问题[118].

近年来, 具有状态约束的非线性系统最优控制问题受到更多的关注, 实现该目标的方法一般称为安全强化学习或者安全 ADP[156−159]. 控制障碍函数是安全控制和最优控制之间的桥梁, 基于控制障碍函数的安全强化学习方法主要分为两类: 状态转换法和惩罚函数法. 状态转换法是先利用控制障碍函数将含有状态约束的系统转换为无约束系统, 然后再利用强化学习或者 ADP 算法求解转换后系统的最优控制策略. 文献 [160] 提出一个“执行−评判−障碍”框架, 利用控制障碍函数将具有状态约束的非零和博弈转换为无约束设计, 有助于在线求解 Nash 均衡问题, 同时保证多玩家连续时间系统的安全性. 文献 [161] 提出一种事件触发的“障碍−执行−评判”方法, 优化了控制器触发和数据传输次数, 同时实现了连续时间系统的安全 H_∞ 控制. 文献 [162] 利用状态转换方法解决了连续时间非线性严格反馈系统的最优输出调节问题. 需要注意的是, 系统转换的前提是模型已知或部分已知. 尽管在利用数据和系统结构特性时能够不要求 $f(x)$ 的信息, 但尚未做到完全的数据驱动. 此外, 转换过程会改变系统原有的优化目标, 从而影响系统的控制性能. 惩罚函

数法是在代价函数中加入控制障碍函数来保证得到的控制策略能够使系统状态满足约束条件. 因此, 不同的控制障碍函数将衍生出不同的代价函数.

本文重点介绍惩罚函数法与控制障碍函数形式. 这里以不等式 $x_{\min}^{[j]} < x^{[j]} < x_{\max}^{[j]}$ 作为基础条件, 其中 $x_{\min}^{[j]}$ 是一个负数, $x_{\max}^{[j]}$ 是一个正数. 对称状态约束是指 $|x_{\min}^{[j]}| = |x_{\max}^{[j]}|$, 而不对称状态约束是指 $|x_{\min}^{[j]}| \neq |x_{\max}^{[j]}|$. 注意, 控制障碍函数的定义在文献 [163–165] 中已详细给出, 这里不再赘述. 接下来, 重点介绍五种控制障碍函数形式:

$$\begin{cases} B_1(x^{[j]}) = \ln\left(\dfrac{x_{\max}^{[j]}}{x_{\max}^{[j]} - x^{[j]}}\right) + \ln\left(\dfrac{x_{\min}^{[j]}}{x_{\min}^{[j]} - x^{[j]}}\right) \\ B_2(x^{[j]}) = -\ln\left(\dfrac{\eta(x_{\max}^{[j]} - x^{[j]})}{\eta(x_{\max}^{[j]} - x^{[j]}) + 1}\right) - \\ \qquad \ln\left(\dfrac{\eta(x^{[j]} - x_{\min}^{[j]})}{\eta(x^{[j]} - x_{\min}^{[j]}) + 1}\right), \qquad \eta > 0 \\ B_3(x^{[j]}) = B_1(x^{[j]}) - \dfrac{x^{[j]}}{x_{\max}^{[j]}} - \dfrac{x^{[j]}}{x_{\min}^{[j]}} \\ B_4(x^{[j]}) = \big(B_2(x^{[j]}) - B_2(0)\big)^2, \qquad \eta = 1 \\ B_5(x^{[j]}) = B_2(x^{[j]}) - B_2(0) - \nabla B_2^{\mathrm{T}}(0)x^{[j]}, \ \eta > 0 \end{cases} \tag{36}$$

其中, $\eta > 0$ 是一个平衡系数. 基于控制障碍函数 $B_1(\cdot)$, 文献 [163] 利用输入输出数据建模的策略迭代算法解决了具有对称状态约束的离散时间仿射系统的最优控制问题. 2023 年, 文献 [164] 使用 $B_1(\cdot)$ 和零和博弈机制, 为具有状态约束的部分不确定非线性离散时间系统提供安全跟踪控制策略. 然而, 对于非对称的状态约束问题, $B_1(\cdot)$ 在约束边界内会出现负值, 这影响了效用函数的正定性. 2021 年, 文献 [165] 提出一种 off-policy 强化学习算法, 利用控制障碍函数 $B_2(\cdot)$ 实现了受状态约束的连续时间非线性系统的数据驱动安全最优控制. 在随机扰动和控制输入矩阵不确定因素的影响下, 2024 年, 文献 [166] 提出一种基于控制障碍函数 $B_2(\cdot)$ 的自适应鲁棒镇定方案, 有效地解决了具有状态约束的不确定非线性系统的非零和微分博弈问题. 文献 [167] 将 $B_2(\cdot)$ 推广到离散时间系统, 通过基于模型的值迭代算法获得安全最优控制策略. 2024 年, 文献 [168] 通过严密的数学推导证明, 在引入具有 $B_2(\cdot)$ 的增广代价函数后, 系统状态能够保持在安全集范围内, 并提出一种 off-policy 安全强化学习算法用于求解安全最优控制策略. 唯一不足的是, $B_2(\cdot)$ 在平衡点处无法取到零, 即 $B_2(0) \neq 0$, 这同样无法保证效用函数的有效性. 相比之下, $B_3(\cdot)$、$B_4(\cdot)$ 以及 $B_5(\cdot)$ 都能够满足在约束边界内大于等于零, 且仅在平衡点处为零, 这符合正定性的条件. 此外, $B_5(\cdot)$ 还具有可调节的优势, 即随着平衡系数的变化, 控制障碍函数的约束功能也会改变. 2025 年, 文献 [169] 采用策略迭代算法和控制障碍函数 $B_3(\cdot)$, 实现了受状态约束和输入约束的离散时间非线性系统的安全最优控制. 2023 年, 文献 [170] 提出控制障碍函数 $B_4(\cdot)$, 并通过在线学习求解值函数, 确保连续时间非线性系统的安全性和最优性. 在 $B_4(\cdot)$ 的基础上, 文献 [171] 解决了具有状态和输入约束的离散时间多层非线性系统的完全合作博弈的安全优化问题. 事实上, 这些针对离散时间系统的安全强化学习算法都是基于模型的, 需要与系统进行交互学习, 而不安全的策略可能会导致后续学习失败. 此外, 已有的增广效用函数通常定义 $U_B(x_k,\ u_k) = x_k^{\mathrm{T}}\mathcal{Q}x_k + u_k^{\mathrm{T}}Ru_k + B(x_k)$, 其中 $B(\cdot)$ 可取式 (36) 中的任意一种形式. 为避免与系统频繁交互, 文献 [172] 在 2024 年提出安全 Q 学习算法, 采用稳定值迭代 Q 学习机制来求解安全最优控制策略, 其 Q 函数的更新过程为:

$$Q_{i+1}(x_k,\ a) = U_B(x_k,\ a) + Q_i\big(x_{k+1},\ u_i(x_{k+1})\big) \tag{37}$$

其中, $U_B(x_k,\ a) = x_{k+1}^{\mathrm{T}}\mathcal{Q}x_{k+1} + B_5(x_{k+1})$. 该效用函数的一个重要特点是在迭代学习过程中将约束施加在 x_{k+1} 上, 这能够保证控制过程中的状态更快地远离约束边界. 实际上, 为兼顾正定性和可调节性, 可以将 $B_2(\cdot)$ 改写为:

$$\bar{B}_2(x^{[j]}) = \ln\left(1 + \frac{\eta}{x_{\max}^{[j]} - x^{[j]}}\right) + \ln\left(1 + \frac{\eta}{x^{[j]} - x_{\min}^{[j]}}\right) \tag{38}$$

这里, $\bar{B}_2(x^{[j]})$ 的形式更简单, 易于神经网络近似, 再利用 $B_5(x^{[j]})$ 的形式进一步提出新的控制障碍函数:

$$B_6(x^{[j]}) = \bar{B}_2(x^{[j]}) - \bar{B}_2(0) - \nabla\bar{B}_2^{\mathrm{T}}(0)x^{[j]} \tag{39}$$

实际上, 控制障碍函数的选取应由实际情况决定. 需要指出的是, 相比于基于模型的安全强化学习算法, 安全 Q 学习算法的优势是能够不与系统交互, 避免学习过程中的不成熟策略损伤系统, 只利用历史数据即可获得安全最优控制策略. 然而, 缺点是无法获取足够有效的位于约束边界内的运行数据. 因此, 将系统模型与安全 Q 学习算法结合, 利用系统模型的预测能力提前产生一些安全数据, 再通过安全 Q 学习算法进行 off-policy 学习是一种更

实用的途径. 需要注意的是, 控制障碍函数包含 $\ln(\cdot)$ 项, 这将增大评判网络对 Q 函数的近似压力, 从而增大迭代算法的计算量和安全最优策略的求解难度. 于是, 结合前述的加速策略评估和策略提升技术, 有助于构建加速的安全 Q 学习算法体系. 此外, 引入深度神经网络作为评判网络, 有利于提高 Q 函数的近似精度.

5.3 博弈 Q 学习算法

零和博弈是指控制输入使得代价函数最小化, 而扰动使得代价函数最大化, 在互相博弈的情况下镇定如下的非线性系统:

$$x_{k+1} = \mathcal{F}(x_k,\ u_k,\ d_k) \tag{40}$$

其中, u_k 为控制向量, d_k 为扰动向量, $\mathcal{F}$ 是未知的系统函数. 假设系统 (40) 在集合 Ω 上是 Lipschitz 连续的且可控, 即至少存在一个能够使得被控系统渐近稳定的控制律.

针对零和博弈问题, 定义无限时域的无折扣代价函数如下所示:

$$J(x_k,\ u_k,\ d_k) = \sum_{\ell=k}^{\infty} U(x_\ell,\ u_\ell,\ d_\ell) \tag{41}$$

其中, $U(x,\ u,\ d) = x^{\mathrm{T}}\mathcal{Q}x + u^{\mathrm{T}}Ru - \delta^2 d^{\mathrm{T}}d$ 为效用函数, $\delta > 0$ 表示扰动衰减水平. 如文献 [173] 中所描述, 设计目标是找到鞍点解 $(u_k^*,\ d_k^*)$ 使得:

$$J(x_k,\ u_k^*,\ d_k) \le J(x_k,\ u_k^*,\ d_k^*) \le J(x_k,\ u_k,\ d_k^*) \tag{42}$$

此外, 鞍点存在的充分条件是:

$$\min_u \max_d J(x_k,\ u_k,\ d_k) = \max_d \min_u J(x_k,\ u_k,\ d_k) \tag{43}$$

根据 Bellman 最优性原理, 最优代价函数满足离散时间 Hamilton-Jacobi-Isaacs 方程:

$$J^*(x_k) = \min_u \max_d \{U(x_k,\ u_k,\ d_k) + J^*(x_{k+1})\} \tag{44}$$

为实现无模型控制, 定义 Q 函数如下所示:

$$\begin{aligned} Q(x_k,\ a,\ b) &= U(x_k,\ a,\ b) + \sum_{\ell=k+1}^{\infty} U(x_\ell,\ u_\ell,\ d_\ell) = \\ &\quad U(x_k,\ a,\ b) + Q(x_{k+1},\ u_{k+1},\ d_{k+1}) = \\ &\quad U(x_k,\ a,\ b) + J(x_{k+1}) \end{aligned} \tag{45}$$

其中, a 和 b 分别表示控制输入和扰动的行为策略[174]. 最优 Q 函数满足:

$$Q^*(x_k,\ a,\ b) = U(x_k,\ a,\ b) + J^*(x_{k+1}) \tag{46}$$

相应地, 最优策略对 $(u^*(x),\ d^*(x))$ 可通过下式获得:

$$\begin{cases} u^*(x) = \arg\min\limits_a Q^*(x,\ a,\ b) \\ d^*(x) = \arg\max\limits_b Q^*(x,\ a,\ b) \end{cases} \tag{47}$$

不难看出, 最优控制策略对 $(u^*(x),\ d^*(x))$ 取决于最优 Q 函数, 但通常无法直接获得 $Q^*(x_k,\ a,\ b)$ 的精确值. 近年来, 面向零和博弈的 Q 学习算法被广泛用于求解模型未知场景下的最优 Q 函数 $Q^*(x_k,\ a,\ b)$[175−178]. 与最优调节、最优跟踪以及约束问题类似, 博弈 Q 学习算法实现的前提是能够获得系统 (40) 的输入输出数据. 令 $D_g = \{x_k^{(l)},\ a^{(l)},\ b^{(l)},\ x_{k+1}^{(l)}\}_{l=1}^{L}$ 表示由多组数据构成的一个数据集, 其中 $x_{k+1}^{(l)} = \mathcal{F}(x_k^{(l)},\ a^{(l)},\ b^{(l)})$.

为解决离散时间非线性零和博弈问题, 文献 [175] 提出一种无模型的全局二次启发式动态规划算法. 其实现原理与调节器中的在线 Q 学习算法类似, 通过将代价函数 $J(x_k,\ u_k,\ d_k)$ 的定义向后设置一步, 能够放宽对系统动力学的要求, 然后构建评判网络用于近似代价函数及其导函数, 构建执行网络和扰动网络分别用于近似控制律以及扰动, 利用在线收集的系统数据持续地更新三个网络权值. 根据迭代形式划分, 博弈 Q 学习算法同样包含值迭代 Q 学习和策略迭代 Q 学习算法. 2024 年, 文献 [176] 提出一种加速的值迭代 Q 学习算法, 根据下式进行 Q 函数更新:

$$\begin{aligned} Q_{i+1}(x_k,\ a,\ b) = \zeta\Big[&U(x_k,\ a,\ b) + \\ &Q_i\big(x_{k+1},\ u_i(x_{k+1}),\ d_i(x_{k+1})\big)\Big] + \\ &(1-\zeta)Q_i(x_k,\ a,\ b) \end{aligned} \tag{48}$$

其中, 可通过式 (49) 的梯度下降法寻找最小化 Q 函数的控制策略和最大化 Q 函数的扰动策略:

$$\begin{cases} u_{i,\,q}(x) = u_{i,\,q-1}(x) - \alpha_c \dfrac{\partial Q_i(x,\ a,\ b)}{\partial a}\Big|_{a=u_{i,\,q-1}(x)} \\ d_{i,\,q}(x) = d_{i,\,q-1}(x) + \alpha_c \dfrac{\partial Q_i(x,\ a,\ b)}{\partial b}\Big|_{b=d_{i,\,q-1}(x)} \end{cases} \tag{49}$$

其中, $u_{i,\,0}(x) = u_{i-1}(x)$; $d_{i,\,0}(x) = d_{i-1}(x)$. 在完成合适的更新次数后, 将最终得到的 $(u_{i,\,q}(x),\ d_{i,\,q}(x))$ 记为提升后的策略对 $(u_i(x),\ d_i(x))$. 实际上, 式 (48) 中的 Q 函数更新过程正是由调节器问题中的加速

值迭代 Q 学习情形推广而来. 此外, 产生数据的行为策略对 $(a,\ b)$ 不是目标策略 $(u_i,\ d_i)$, 因此该算法属于 off-policy 学习. 接下来, 针对非线性零和博弈问题, 我们给出基于 off-policy 学习的策略迭代 Q 学习算法, 其 Q 函数更新过程为:

$$Q_i(x_k,\ a,\ b) = U(x_k,\ a,\ b) + Q_i\big(x_{k+1},\ u_i(x_{k+1}),\ d_i(x_{k+1})\big) \tag{50}$$

与调节器的策略迭代算法一致, 这里需要一个初始容许的控制策略和扰动策略. 2022 年, 文献 [177] 将事件触发机制引入式 (50) 中, 从而节省了计算和通信资源. 2023 年, 文献 [178] 构建了基于 on-policy 学习的策略迭代 Q 学习算法:

$$\begin{aligned} Q_i(x_k,\ u_i(x_k),\ d_i(x_k)) = \\ U\big(x_k,\ u_i(x_k),\ d_i(x_k)\big) + \\ Q_i\big(x_{k+1},\ u_i(x_{k+1}),\ d_i(x_{k+1})\big) \end{aligned} \tag{51}$$

然后引入经验回放机制, 以 off-policy 学习形式获取最优策略对. 需要指出, 几乎所有与零和博弈相关的效用函数都如式 (41) 中所示, 这需要确定 δ 值. 为简化数据驱动设计过程, 本文构建一种新的效用函数形式为 $U(x_k,\ a,\ b) = x_{k+1}^{\mathrm{T}}\mathcal{Q}x_{k+1}$, 这有效保证了学习过程中效用函数的正定性. 此外, 该形式也能够方便地推广到零和博弈的跟踪控制和安全控制, 改进的效用函数形式分别为:

$$\begin{cases} U(e_k,\ r_k,\ a,\ b) = e_{k+1}^{\mathrm{T}}\mathcal{Q}e_{k+1} \\ U_B(x_k,\ a,\ b) = x_{k+1}^{\mathrm{T}}\mathcal{Q}x_{k+1} + B_6(x_{k+1}) \end{cases} \tag{52}$$

应该看到, 面向调节器的性能改进 Q 学习机制, 都能够平移推广到零和博弈问题, 充分证明了加速、迁移、集成等 Q 学习算法的有效性和通用性.

6 污水处理系统典型应用

污水处理过程是一个涉及物理、化学和生物交互反应的复杂动态系统, 融合人类行为与决策, 呈现出多要素性、强非线性、非平稳性等特点, 实现污水处理过程的“稳定−安全−经济”一体化智慧运行是一项具有挑战性的任务. 污水处理过程的基本控制目标是, 在具有扰动和变量约束的情况下, 确保溶解氧浓度和硝态氮浓度跟踪上设定值[179]. 因此, 污水处理过程包含本文所阐述的跟踪控制、安全控制、鲁棒控制问题, 这有利于总结数据驱动的 ADP 算法在实际工业系统中的应用效果和存在问题.

本文探讨污水处理应用中的三类问题. 污水处理过程涉及诸多环节, 系统模型难以建立, 如何实现跟踪控制是需要解决的第一个科学问题. 2022 年, 文献 [180] 使用在线 Q 学习算法解决了污水处理过程中溶解氧和硝态氮浓度的跟踪控制问题. 随后, 文献 [181] 将基于经验回放的策略梯度 ADP 算法用于解决污水处理过程中的溶解氧和硝态氮浓度跟踪问题. 文献 [182] 提出面向污水处理系统的离线强化学习算法, 本质是采用经典控制器与系统交互产生数据, 然后通过策略梯度 ADP 方法获得最优控制策略. 2024 年, 文献 [116] 提出融合知识迁移的在线 Q 学习算法, 通过利用与代价函数相关的先验知识, 获得了更优越的跟踪性能. 随后, 文献 [117] 进一步提出补充 Q 学习算法, 通过引入经验回放机制和经典控制器, 有效降低了试错成本和控制器的设计复杂度. 通过对比实验结果, 作者发现利用迁移学习和补充学习的方法获得了更好的控制性能[116−117], 这主要是因为污水处理过程中控制变量的数值较大, 例如与溶解氧浓度相关的控制变量范围为 [0, 240], 而与硝态氮浓度相关的控制变量范围为 [0, 92 230]. 因此, 仅仅依靠执行网络输出的控制律难以直接满足控制需求. 从这个角度看, 设计补充控制器或者前馈控制器是必要的. 此外, 当控制变量特别大时, 在线学习的执行网络和评判网络难以实现有效的近似, 而将控制变量的更新方式改进为增量形式, 能够大幅降低神经网络的近似压力. 当前, 已有一些基于神经网络建模的增量 ADP 方法[183−185], 但主要是将非线性系统转换为线性系统进行增量控制器的设计. 目前, 尚没有面向非线性系统的数据驱动增量 Q 学习算法. 此外, 由于增量 Q 学习控制与增量式 PID 控制具有相同形式的策略, 更利于实际的工业系统应用. 污水处理过程运行工况复杂, 受物理限制和各种变量约束, 如何实现安全控制是需要解决的第二个科学问题. 对于受控制约束的一般非线性系统, 约束范围通常为一个正值和一个负值, 而污水处理过程中控制约束的范围都是正值, 这要求设计新的约束函数. 此外, 溶解氧和硝态氮浓度的跟踪误差需要限制在一定范围内, 否则会引起出水质量下降, 这实际上是一个状态约束问题. 传统的安全强化学习方法由于参数固定, 难以应对时变的大规模场景. 通过开展具有自适应能力的安全强化学习研究, 不断地调整和改善污水处理过程的设计策略, 从而保证系统的安全和高效运行是值得关注的重要方向. 污水处理过程具有非平稳性的特点, 扰动因素不明确, 如何实现鲁棒控制是需要解决的第三个科学问题. 2024 年, 文献 [186] 提出一种数据驱动的鲁棒 ADP 算法, 设计一个鲁棒项来抑制污水处理系统和环境的未知干扰, 有效地平衡了系统能耗和控制性能. 然而, 鲁棒 ADP 算

法的控制性能有待提升, 设计更简单有效的鲁棒补偿器对于实际工业系统控制具有重要意义[187]. 作为数据驱动自适应评判的典型应用场景, 我们相信, 污水处理系统的智慧运行方面将会有更多的有益成果.

7 结束语

本文总结了数据驱动自适应评判方法的基本原理和研究现状, 并提出一系列先进的评判学习机制和函数设计形式. 首先, 从非线性系统的最优调节问题入手, 阐述在线 Q 学习算法和迭代 Q 学习算法的实现结构、理论特性以及数据学习形式; 其次, 为实现高效学习, 回顾了已有的加速 Q 学习和迁移 Q 学习算法, 并设计了改进的加速和迁移学习机制; 再次, 阐述 Q 学习算法在解决跟踪、安全、博弈等问题上的研究成果, 并构建了改进的效用函数形式; 最后, 分析了数据驱动自适应评判方法在污水处理过程中的应用和挑战, 重点指出进一步开展增量、安全、鲁棒 Q 学习算法研究具有实际意义. 随着数据驱动自适应评判控制的理论研究日渐深入, 利用强化学习、迁移学习、深度学习等人工智能技术创新作为突破口, 积极推动自动化与智能系统产业技术发展, 具有重要的应用前景.

References

1 Zhang Hua-Guang, Zhang Xin, Luo Yan-Hong, Yang Jun. An overview of research on adaptive dynamic programming. *Acta Automatica Sinica*, 2013, **39**(4): 303−311
(张化光, 张欣, 罗艳红, 杨珺. 自适应动态规划综述. 自动化学报, 2013, **39**(4): 303−311)

2 Lewis F L, Vrabie D, Vamvoudakis K G. Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers. *IEEE Control Systems Magazine*, 2012, **32**(6): 76−105

3 Sutton R S, Barto A G. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 1998.

4 Werbos P J. *Approximate Dynamic Programming for Real-time Control and Neural Modeling*. New York: Van Nostrand Reinhold, 1992.

5 Liu De-Rong, Li Hong-Liang, Wang Ding. Data-based self-learning optimal control: Research progress and prospects. *Acta Automatica Sinica*, 2013, **39**(11): 1858−1870
(刘德荣, 李宏亮, 王鼎. 基于数据的自学习优化控制: 研究进展与展望. 自动化学报, 2013, **39**(11): 1858−1870)

6 Mao R Q, Cui R X, Chen C L P. Broad learning with reinforcement learning signal feedback: Theory and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 2022, **33**(7): 2952−2964

7 Shi Y X, Hu Q L, Li D Y, Lv M L. Adaptive optimal tracking control for spacecraft formation flying with event-triggered input. *IEEE Transactions on Industrial Informatics*, 2023, **19**(5): 6418−6428

8 Sun Chang-Yin, Mu Chao-Xu. Important scientific problems of multi-agent deep reinforcement learning. *Acta Automatica Sinica*, 2020, **46**(7): 1301−1312
(孙长银, 穆朝絮. 多智能体深度强化学习的若干关键科学问题. 自动化学报, 2020, **46**(7): 1301−1312)

9 Wei Q L, Liao Z H, Shi G. Generalized actor-critic learning optimal control in smart home energy management. *IEEE Transactions on Industrial Informatics*, 2021, **17**(10): 6614−6623

10 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)

11 Sun J Y, Dai J, Zhang H G, Yu S H, Xu S, Wang J J. Neural-network-based immune optimization regulation using adaptive dynamic programming. *IEEE Transactions on Cybernetics*, 2023, **53**(3): 1944−1953

12 Liu D R, Ha M M, Xue S. State of the art of adaptive dynamic programming and reinforcement learning. *CAAI Artificial Intelligence Research*, 2022, **1**(2): 93−110

13 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

14 Wang Ding, Zhao Ming-Ming, Ha Ming-Ming, Ren Jin. *Intelligent Control and Reinforcement Learning: Advanced Value Iteration Critic Design*. Beijing: Posts and Telecommunications Press, 2024.
(王鼎, 赵明明, 哈明鸣, 任进. 智能控制与强化学习: 先进值迭代评判设计. 北京: 人民邮电出版社, 2024.)

15 Sun Jing-Liang, Liu Chun-Sheng. An overview on the adaptive dynamic programming based missile guidance law. *Acta Automatica Sinica*, 2017, **43**(7): 1101−1113
(孙景亮, 刘春生. 基于自适应动态规划的导弹制导律研究综述. 自动化学报, 2017, **43**(7): 1101−1113)

16 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**: 12315−12346

17 Wang D, Ha M M, Zhao M M. The intelligent critic framework for advanced optimal control. *Artificial Intelligence Review*, 2022, **55**(1): 1−22

18 Al-Tamimi A, Lewis F L, Abu-Khalaf M. Discrete-time nonlinear HJB solution using approximate dynamic programming: Convergence proof. *IEEE Transactions on Systems, Man, and Cybernetics-Part B: Cybernetics*, 2008, **38**(4): 943−949

19 Li H L, Liu D R. Optimal control for discrete-time affine nonlinear systems using general value iteration. *IET Control Theory and Applications*, 2012, **6**(18): 2725−2736

20 Wei Q L, Liu D R, Lin H Q. Value iteration adaptive dynamic programming for optimal control of discrete-time nonlinear systems. *IEEE Transactions on Cybernetics*, 2016, **46**(3): 840−853

21 Wang D, Zhao M M, Ha M M, Qiao J F. Stability and admissibility analysis for zero-sum games under general value iteration formulation. *IEEE Transactions on Neural Networks and Learning Systems*, 2023, **34**(11): 8707−8718

22 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

23 Heydari A. Stability analysis of optimal adaptive control under value iteration using a stabilizing initial policy. *IEEE Transactions on Neural Networks and Learning Syetems*, 2018, **29**(9): 4522−4527

24 Wei Q L, Lewis F L, Liu D R, Song R Z, Lin H Q. Discrete-time local value iteration adaptive dynamic programming: Convergence analysis. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2018, **48**(6): 875−891

25 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

26 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

27 Luo B, Liu D R, Huang T W, Yang X, Ma H W. Multi-step heuristic dynamic programming for optimal control of nonlinear discrete-time systems. *Information Sciences*, 2017, **411**: 66−83

28 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

29 Rao J, Wang J C, Xu J H, Zhao S W. Optimal control of nonlinear system based on deterministic policy gradient with eligibility traces. *Nonlinear Dynamics*, 2023, **111**: 20041−20053

30 Yu L Y, Liu W B, Liu Y R, Alsaadi F E. Learning-based T-sHDP (λ) for optimal control of a class of nonlinear discrete-time systems. *International Journal of Robust and Nonlinear Control*, 2022, **32**(5): 2624−2643

31 Al-Dabooni S, Wunsch D. An improved n-step value gradient learning adaptive dynamic programming algorithm for online learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2020, **31**(4): 1155−1169

32 Wang J Y, Wang D, Li X, Qiao J F. Dichotomy value iteration with parallel learning design towards discrete-time zero-sum games. *Neural Networks*, 2023, **167**: 751−762

33 Wei Q L, Wang L X, Lu J W, Wang F Y. Discrete-time self-learning parallel control. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2022, **52**(1): 192−204

34 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

35 Ha M M, Wang D, Liu D R. Novel discounted adaptive critic control designs with accelerated learning formulation. *IEEE Transactions on Cybernetics*, 2024, **54**(5): 3003−3016

36 Wang D, Huang H M, Liu D R, Zhao M M, Qiao J F. Evolution-guided adaptive dynamic programming for nonlinear optimal control. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2024, **54**(10): 6043−6054

37 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

38 Liu D R, Wei Q L. Generalized policy iteration adaptive dynamic programming for discrete-time nonlinear systems. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2015, **45**(12): 1577−1591

39 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

40 Luo B, Yang Y, Wu H N, Huang T W. Balancing value iteration and policy iteration for discrete-time control. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2020, **50**(11): 3948−3958

41 Li T, Wei Q L, Wang F Y. Multistep look-ahead policy iteration for optimal control of discrete-time nonlinear systems with isoperimetric constraints. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2024, **54**(3): 1414−1426

42 Yang Y L, Kiumarsi B, Modares H, Xu C Z. Model-free λ-policy iteration for discrete-time linear quadratic regulation. *IEEE Transactions on Neural Networks and Learning Systems*, 2023, **34**(2): 635−649

43 Huang H M, Wang D, Wang H, Wu J L, Zhao M M. Novel generalized policy iteration for efficient evolving control of nonlinear systems. *Neurocomputing*, 2024, **608**: Article No. 128418

44 Dierks T, Jagannathan S. Online optimal control of affine nonlinear discrete-time systems with unknown internal dynamics by using time-based policy update. *IEEE Transactions on Neural Networks and Learning Systems*, 2012, **23**(7): 1118−1129

45 Wang D, Xin P, Zhao M M, Qiao J F. Intelligent optimal control of constrained nonlinear systems via receding-horizon heuristic dynamic programming. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2024, **54**(1): 287−299

46 Moghadam R, Natarajan P, Jagannathan S. Online optimal adaptive control of partially uncertain nonlinear discrete-time systems using multilayer neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2022, **33**(9): 4840−4850

47 Zhang H G, Qin C B, Jiang B, Luo Y H. Online adaptive policy learning algorithm for H_∞ state feedback control of unknown affine nonlinear discrete-time systems. *IEEE Transactions on Cybernetics*, 2014, **44**(12): 2706−2718

48 Ming Z Y, Zhang H G, Yan Y Q, Zhang J. Tracking control of discrete-time system with dynamic event-based adaptive dynamic programming. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 2022, **69**(8): 3570−3574

49 Luo Biao, Ouyang Zhi-Hua, Yi Xin-Ning, Liu De-Rong. Adaptive dynamic programming based visual servoing tracking control for mobile robots. *Acta Automatica Sinica*, 2023, **49**(11): 2286−2296
(罗彪, 欧阳志华, 易昕宁, 刘德荣. 基于自适应动态规划的移动机器人视觉伺服跟踪控制. 自动化学报, 2023, **49**(11): 2286−2296)

50 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

51 Dong L, Zhong X N, Sun C Y, He H B. Adaptive event-triggered control based on heuristic dynamic programming for nonlinear discrete-time systems. *IEEE Transactions on Neural Networks and Learning Systems*, 2017, **28**(7): 1594−1605

52 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

53 Yang X, Wang D. Reinforcement learning for robust dynamic event-driven constrained control. *IEEE Transactions on Neural Networks and Learning Systems*, DOI: 10.1109/TNNLS.2024.3394251

54 Wang Ding. Research progress on learning-based robust adaptive critic control. *Acta Automatica Sinica*, 2019, **45**(6): 1031−1043
(王鼎. 基于学习的鲁棒自适应评判控制研究进展. 自动化学报, 2019, **45**(6): 1031−1043)

55 Ren H, Jiang B, Ma Y J. Zero-sum differential game-based fault-tolerant control for a class of affine nonlinear systems. *IEEE Transactions on Cybernetics*, 2024, **54**(2): 1272−1282

56 Zhang S C, Zhao B, Liu D R, Zhang Y W. Event-triggered decentralized integral sliding mode control for input-constrained nonlinear large-scale systems with actuator failures. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2024, **54**(3): 1914−1925

57 Wei Q L, Zhu L, Song R Z, Zhang P J, Liu D R, Xiao J. Model-free adaptive optimal control for unknown nonlinear multiplayer nonzero-sum game. *IEEE Transactions on Neural Networks and Learning Systems*, 2022, **33**(2): 879−892

58 Ye J, Bian Y G, Luo B, Hu M J, Xu B, Ding R. Costate-supplement ADP for model-free optimal control of discrete-time nonlinear systems. *IEEE Transactions on Neural Networks and Learning Systems*, 2024, **35**(1): 45−59

59 Li Y Q, Yang C Z, Hou Z S, Feng Y J, Yin C K. Data-driven approximate Q-learning stabilization with optimality error

bound analysis. *Automatica*, 2019, **103**: 435−442

60 Al-Dabooni S, Wunsch D C. Online model-free *n*-step HDP with stability analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 2020, **31**(4): 1255−1269

61 Ni Z, He H B, Zhong X N, Prokhorov D V. Model-free dual heuristic dynamic programming. *IEEE Transactions on Neural Networks and Learning Systems*, 2015, **26**(8): 1834−1839

62 Wang D, Ha M M, Qiao J F. Self-learning optimal regulation for discrete-time nonlinear systems under event-driven formulation. *IEEE Transactions on Automatic Control*, 2020, **65**(3): 1272−1279

63 Wang D, Ha M M, Qiao J F. Data-driven iterative adaptive critic control toward an urban wastewater treatment plant. *IEEE Transactions on Industrial Electronics*, 2021, **68**(8): 7362−7369

64 Wang D, Hu L Z, Zhao M M, Qiao J F. Adaptive critic for event-triggered unknown nonlinear optimal tracking design with wastewater treatment applications. *IEEE Transactions on Neural Networks and Learning Systems*, 2023, **34**(9): 6276−6288

65 Zhu L, Wei Q L, Guo P. Synergetic learning neuro-control for unknown affine nonlinear systems with asymptotic stability guarantees. *IEEE Transactions on Neural Networks and Learning Systems*, 2025, **36**(2): 3479−3489

66 Pang B, Jiang Z P. Adaptive optimal control of linear periodic systems: An off-policy value iteration approach. *IEEE Transactions on Automatic Control*, 2021, **66**(2): 888−894

67 Xu Y S, Zhao Z G, Yin S. Performance optimization and fault-tolerance of highly dynamic systems via Q-learning with an incrementally attached controller gain system. *IEEE Transactions on Neural Networks and Learning Systems*, 2023, **34**(11): 9128−9138

68 Yang X, Xu M M, Wei Q L. Adaptive dynamic programming for nonlinear-constrained H_∞ control. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2023, **53**(7): 4393−4403

69 Werbos P J. Neural networks for control and system identification. In: Proceedings of the 28th IEEE Conference on Decision and Control. Tampa, FL, USA: IEEE, 1989. 260−265

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

71 Watkins C. Learning From Delayed Rewards [Ph.D. dissertation], King′s College of Cambridge, UK, 1989.

72 Al-Tamimi A, Lewis F L, Abu-Khalaf M. Model-free Q-learning designs for linear discrete-time zero-sum games with application to H-infinity control. *Automatica*, 2007, **43**(3): 473−481

73 Kiumarsi B, Lewis F L, Modares H, Karimpour A, Naghibi-Sistani M. Reinforcement Q-learning for optimal tracking control of linear discrete-time systems with unknown dynamics. *Automatica*, 2014, **50**(4): 1167−1175

74 Jiang Y, Jiang Z P. Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics. *Automatica*, 2012, **48**(10): 2699−2704

75 Kiumarsi B, Lewis F L, Jiang Z P. H_∞ control of linear discrete-time systems: Off-policy reinforcement learning. *Automatica*, 2017, **78**: 144−152

76 Farjadnasab M, Babazadeh M. Model-free LQR design by Q-function learning. *Automatica*, 2022, **137**: Article No. 110060

77 Lopez V G, Alsalti M, Müller M A. Efficient off-policy Q-learning for data-based discrete-time LQR problems. *IEEE Transactions on Automatic Control*, 2023, **68**(5): 2922−2933

78 Nguyen H, Dang H B, Dao P N. On-policy and off-policy Q-learning strategies for spacecraft systems: An approach for time-varying discrete-time without controllability assumption of augmented system. *Aerospace Science and Technology*, 2024, **146**: Article No. 108972

79 Skach J, Kiumarsi B, Lewis F L, Straka O. Actor-critic off-policy learning for optimal control of multiple-model discrete-time systems. *IEEE Transactions on Cybernetics*, 2018, **48**(1): 29−40

80 Wen Y L, Zhang H G, Ren H, Zhang K. Off-policy based adaptive dynamic programming method for nonzero-sum games on discrete-time system. *Journal of the Franklin Institute*, 2020, **357**(12): 8059−8081

81 Xu Y, Wu Z G. Data-efficient off-policy learning for distributed optimal tracking control of HMAS with unidentified exosystem dynamics. *IEEE Transactions on Neural Networks and Learning Systems*, 2024, **35**(3): 3181−3190

82 Cui L L, Pang B, Jiang Z P. Learning-based adaptive optimal control of linear time-delay systems: A policy iteration approach. *IEEE Transactions on Automatic Control*, 2024, **69**(1): 629−636

83 Amirparast A, Sani S K H. Off-policy reinforcement learning algorithm for robust optimal control of uncertain nonlinear systems. *International Journal of Robust and Nonlinear Control*, 2024, **34**(8): 5419−5437

84 Qasem O, Gao W N, Vamvoudakis K G. Adaptive optimal control of continuous-time nonlinear affine systems via hybrid iteration. *Automatica*, 2023, **157**: Article No. 111261

85 Jiang H Y, Zhou B, Duan G R. Modified λ-policy iteration based adaptive dynamic programming for unknown discrete-time linear systems. *IEEE Transactions on Neural Networks and Learning Systems*, 2024, **35**(3): 3291−3301

86 Zhao J G, Yang C Y, Gao W N, Park J H. Novel single-loop policy iteration for linear zero-sum games. *Automatica*, 2024, **163**: Article No. 111551

87 Xiao Zhen-Fei, Li Jin-Na. Two-player optimization control based on off-policy Q-learning algorithm. *Control Engineering of China*, 2022, **29**(10): 1874−1880
(肖振飞, 李金娜. 基于非策略 Q 学习方法的两个个体优化控制. 控制工程, 2022, **29**(10): 1874−1880)

88 Liu Y, Zhang H G, Yu R, Xing Z X. H_∞ tracking control of discrete-time system with delays via data-based adaptive dynamic programming. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2020, **50**(11): 4078−4085

89 Zhang H G, Liu Y, Xiao G Y, Jiang H. Data-based adaptive dynamic programming for a class of discrete-time systems with multiple delays. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2020, **50**(2): 432−441

90 Tan X F, Li Y, Liu Y. Stochastic linear quadratic optimal tracking control for discrete-time systems with delays based on Q-learning algorithm. *AIMS Mathematics*, 2023, **8**(5): 10249−10265

91 Zhang L L, Zhang H G, Sun J Y, Yue X. ADP-based fault-tolerant control for multiagent systems with semi-markovian jump parameters. *IEEE Transactions on Cybernetics*, 2024, **54**(10): 5952−5962

92 Li Y, Zhang H, Wang Z P, Huang C, Yan H C. Data-driven decentralized control for large-scale systems with sparsity and communication delays. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2023, **53**(9): 5614−5624

93 Shen X Y, Li X J. Data-driven output-feedback LQ secure control for unknown cyber-physical systems against sparse actuator attacks. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2021, **51**(9): 5708−5720

94 Qasem O, Davari M, Gao W N, Kirk D R, Chai T Y. Hybrid iteration ADP algorithm to solve cooperative, optimal output regulation problem for continuous-time, linear, multiagent systems: Theory and application in islanded modern microgrids with IBRs. *IEEE Transactions on Industrial Electronics*, 2024, **71**(1): 834−845

95 Zhang H G, Liang H J, Wang Z S, Feng T. Optimal output regulation for heterogeneous multiagent systems via adaptive dynamic programming. *IEEE Transactions on Neural Networks and Learning Systems*, 2017, **28**(1): 18−29

96 Wang W, Chen X. Model-free optimal containment control of multi-agent systems based on actor-critic framework. *Neurocomputing*, 2018, **314**(7): 242−250

97 Cui L L, Wang S, Zhang J F, Zhang D S, Lai J, Zheng Y, et al. Learning-based balance control of wheel-legged robots. *IEEE Robotics and Automation Letters*, 2021, **6**(4): 7667−7674

98 Liu T, Cui L L, Pang B, Jiang Z P. A unified framework for data-driven optimal control of connected vehicles in mixed traffic. *IEEE Transactions on Intelligent Vehicles*, 2023, **8**(8): 4131−4145

99 Davari M, Gao W N, Aghazadeh A, Blaabjerg F, Lewis F L. An optimal synchronization control method of PLL utilizing adaptive dynamic programming to synchronize inverter-based resources with unbalanced, low-inertia, and very weak grids. *IEEE Transactions on Automation Science and Engineering*, 2025, **22**: 24−42

100 Wang Z Y, Wang Y Q, Davari M, Blaabjerg F. An effective PQ-decoupling control scheme using adaptive dynamic programming approach to reducing oscillations of virtual synchronous generators for grid connection with different impedance types. *IEEE Transactions on Industrial Electronics*, 2024, **71**(4): 3763−3775

101 Si J, Wang Y T. Online learning control by association and reinforcement. *IEEE Transactions on Neural Networks*, 2001, **12**(2): 264−276

102 Liu F, Sun J, Si J, Guo W T, Mei S W. A boundedness result for the direct heuristic dynamic programming. *Neural Networks*, 2012, **32**: 229−235

103 Sokolov Y, Kozma R, Werbos L D, Werbos P J. Complete stability analysis of a heuristic approximate dynamic programming control design. *Automatica*, 2015, **59**: 9−18

104 Malla N, Ni Z. A new history experience replay design for model-free adaptive dynamic programming. *Neurocomputing*, 2017, **266**(29): 141−149

105 Luo B, Wu H N, Huang T W, Liu D R. Data-based approximate policy iteration for affine nonlinear continuous-time optimal control design. *Automatica*, 2014, **50**(12): 3281−3290

106 Zhao D B, Xia Z P, Wang D. Model-free optimal control for affine nonlinear systems with convergence analysis. *IEEE Transactions on Automation Science and Engineering*, 2015, **12**(4): 1461−1468

107 Xu J H, Wang J C, Rao J, Zhong Y J, Wu S Y, Sun Q F. Parallel cross entropy policy gradient adaptive dynamic programming for optimal tracking control of discrete-time nonlinear systems. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2024, **54**(6): 3809−3821

108 Wei Q L, Lewis F L, Sun Q Y, Yan P F, Song R Z. Discrete-time deterministic Q-learning: A novel convergence analysis. *IEEE Transactions on Cybernetics*, 2017, **47**(5): 1224−1237

109 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)

110 Qiao J F, Zhao M M, Wang D, Ha M M. Adjustable iterative Q-learning schemes for model-free optimal tracking control. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2024, **54**(2): 1202−1213

111 Ni Z, Malla N, Zhong X N. Prioritizing useful experience replay for heuristic dynamic programming-based learning systems. *IEEE Transactions on Cybernetics*, 2019, **49**(11): 3911−3922

112 Al-Dabooni S, Wunsch D. The boundedness conditions for model-free HDP (λ). *IEEE Transactions on Neural Networks and Learning Systems*, 2019, **30**(7): 1928−1942

113 Zhao Q T, Si J, Sun J. Online reinforcement learning control by direct heuristic dynamic programming: From time-driven to event-driven. *IEEE Transactions on Neural Networks and Learning Systems*, 2022, **33**(8): 4139−4144

114 Wei Q L, Liao Z H, Song R Z, Zhang P J, Wang Z, Xiao J. Self-learning optimal control for ice-storage air conditioning systems via data-based adaptive dynamic programming. *IEEE Transactions on Industrial Electronics*, 2021, **68**(4): 3599−3608

115 Zhao J, Wang T Y, Pedrycz W, Wang W. Granular prediction and dynamic scheduling based on adaptive dynamic programming for the blast furnace gas system. *IEEE Transactions on Cybernetics*, 2021, **51**(4): 2201−2214

116 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*, 2024, **20**(2): 1488−1497

117 Qiao J F, Zhao M M, Wang D, Li M H. Action-dependent heuristic dynamic programming with experience replay for wastewater treatment processes. *IEEE Transactions on Industrial Informatics*, 2024, **20**(4): 6257−6265

118 Luo B, Liu D R, Wu H N. Adaptive constrained optimal control design for data-based nonlinear discrete-time systems with critic-only structure. *IEEE Transactions on Neural Networks and Learning Systems*, 2018, **29**(6): 2099−2111

119 Zhao M M, Wang D, Qiao J F. Stabilizing value iteration Q-learning for online evolving control of discrete-time nonlinear systems. *Nonlinear Dynamics*, 2024, **112**: 9137−9153

120 Xiang Z R, Li P C, Zou W C, Ahn C K. Data-based optimal switching and control with admissibility guaranteed Q-learning. *IEEE Transactions on Neural Networks and Learning Systems*, DOI: 10.1109/TNNLS.2024.3405739

121 Li X F, Dong L, Xue L, Sun C Y. Hybrid reinforcement learning for optimal control of non-linear switching system. *IEEE Transactions on Neural Networks and Learning Systems*, 2023, **34**(11): 9161−9170

122 Li J N, Chai T Y, Lewis F L, Ding Z T, Jiang Y. Off-policy interleaved Q-learning: Optimal control for affine nonlinear discrete-time systems. *IEEE Transactions on Neural Networks and Learning Systems*, 2019, **30**(5): 1308−1320

123 Song S J, Zhao M M, Gong D W, Zhu M L. Convergence and stability analysis of value iteration Q-learning under non-discounted cost for discrete-time optimal control. *Neurocomputing*, 2024, **606**: Article No. 128370

124 Song S J, Zhu M L, Dai X L, Gong D W. Model-free optimal tracking control of nonlinear input-affine discrete-time systems via an iterative deterministic Q-learning algorithm. *IEEE Transactions on Neural Networks and Learning Systems*, 2024, **35**(1): 999−1012

125 Wei Q L, Liu D R. A novel policy iteration based deterministic Q-learning for discrete-time nonlinear systems. *Science China Information Sciences*, 2015, **58**(12): 1−15

126 Yan P F, Wang D, Li H L, Liu D R. Error bound analysis of Q-function for discounted optimal control problems with policy iteration. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2017, **47**(7): 1207−1216

127 Wang W, Chen X, Fu H, Wu M. Model-free distributed consensus control based on actor-critic framework for discrete-time nonlinear multiagent systems. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2020, **50**(11): 4123−4134

128 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

129 Zhang Y W, Zhao B, Liu D R. Deterministic policy gradient adaptive dynamic programming for model-free optimal control. *Neurocomputing*, 2020, **387**: 40−50

130 Xu J H, Wang J C, Rao J, Zhong Y J, Zhao S W. Twin deterministic policy gradient adaptive dynamic programming for optimal control of affine nonlinear discrete-time systems. *International Journal of Control, Automation, and Systems*, 2022, **20**(9): 3098−3109

131 Xu J H, Wang J C, Rao J, Wu S Y, Zhong Y J. Adaptive dynamic programming for optimal control of discrete-time nonlinear systems with trajectory-based initial control policy. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2024, **54**(3): 1489−1501

132 Lin M D, Zhao B. Policy optimization adaptive dynamic programming for optimal control of input-affine discrete-time nonlinear systems. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2023, **53**(7): 4339−4350

133 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

134 Luo B, Yang Y, Liu D R. Adaptive Q-learning for data-based optimal output regulation with experience replay. *IEEE Transactions on Cybernetics*, 2018, **48**(12): 3337−3348

135 Qasem O, Gutierrez H, Gao W N. Experimental validation of data-driven adaptive optimal control for continuous-time systems via hybrid iteration: An application to rotary inverted pendulum. *IEEE Transactions on Industrial Electronics*, 2024, **71**(6): 6210−6220

136 Li Man-Yuan, Luo Fei, Gu Chun-Hua, Luo Yong-Jun, Ding Wei-Chao. Adams algorithm based on adaptive momentum update strategy. *Journal of University of Shanghai for Science and Technology*, 2023, **45**(2): 112−119
(李满园, 罗飞, 顾春华, 罗勇军, 丁炜超. 基于自适应动量更新策略的 Adams 算法. 上海理工大学学报, 2023, **45**(2): 112−119)

137 Jiang Zhi-Xia, Song Jia-Shuai, Liu Yu-Ning. An improved adaptive momentum gradient descent algorithm. *Journal of Huazhong University of Science and Technology (Natural Science Edition)*, 2023, **51**(5): 137−143
(姜志侠, 宋佳帅, 刘宇宁. 一种改进的自适应动量梯度下降算法. 华中科技大学学报 (自然科学版), 2023, **51**(5): 137−143)

138 Jiang Wen-Han, Jiang Zhi-Xia, Sun Xue-Lian. A gradient descent algorithm with modified learning rate. *Journal of Changchun University of Science and Technology (Natural Science Edition)*, 2023, **46**(6): 112−120
(姜文翰, 姜志侠, 孙雪莲. 一种修正学习率的梯度下降算法. 长春理工大学学报 (自然科学版), 2023, **46**(6): 112−120)

139 Zhao B, Shi G, Liu D R. Event-triggered local control for nonlinear interconnected systems through particle swarm optimization-based adaptive dynamic programming. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2023, **53**(12): 7342−7353

140 Zhang L J, Zhang K, Xie X P, Chadli M. Adaptive critic control with knowledge transfer for uncertain nonlinear dynamical dystems: A reinforcement learning approach. *IEEE Transactions on Automation Science and Engineering*, DOI: 10.1109/TASE.2024.3453926

141 Gao X, Si J, Huang H. Reinforcement learning control with knowledge shaping. *IEEE Transactions on Neural Networks and Learning Systems*, 2024, **35**(3): 3156−3167

142 Gao X, Si J, Wen Y, Li M H, Huang H. Reinforcement learning control of robotic knee with human-in-the-loop by flexible policy iteration. *IEEE Transactions on Neural Networks and Learning Systems*, 2022, **33**(10): 5873−5887

143 Guo W T, Liu F, Si J, He D W, Harley R, Mei S W. Online supplementary ADP learning controller design and application to power system frequency control with large-scale wind energy integration. *IEEE Transactions on Neural Networks and Learning Systems*, 2016, **27**(8): 1748−1761

144 Zhao M M, Wang D, Ren J, Qiao J. Integrated online Q-learning design for wastewater treatment processes. *IEEE Transactions on Industrial Informatics*, 2025, **21**(2): 1833−1842

145 Zhang H G, Wei Q L, Luo Y H. A novel infinite-time optimal tracking control scheme for a class of discrete-time nonlinear systems via the greedy HDP iteration algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 2008, **38**(4): 937−942

146 Song S J, Gong D W, Zhu M L, Zhao Y Y, Huang C. Data-driven optimal tracking control for discrete-time nonlinear systems with unknown dynamics using deterministic ADP. *IEEE Transactions on Neural Networks and Learning Systems*, 2025, **36**(1): 1184−1198

147 Luo B, Liu D R, Huang T W, Wang D. Model-free optimal tracking control via critic-only Q-learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2016, **27**(10): 2134−2144

148 Li C, Ding J L, Lewis F L, Chai T Y. A novel adaptive dynamic programming based on tracking error for nonlinear discrete-time systems. *Automatica*, 2021, **129**: Article No. 109687

149 Wang D, Gao N, Ha M M, Zhao M M, Wu J L, Qiao J F. Intelligent-critic-based tracking control of discrete-time input-affine systems and approximation error analysis with application verification. *IEEE Transactions on Cybernetics*, 2024, **54**(8): 4690−4701

150 Liang Z T, Ha M M, Liu D R, Wang Y H. Stable approximate Q-learning under discounted cost for data-based adaptive tracking control. *Neurocomputing*, 2024, **568**: Article No. 127048

151 Wang Y, Wang D, Zhao M M, Liu A, Qiao J F. Adjustable iterative Q-learning for advanced neural tracking control with stability guarantee. *Neurocomputing*, 2024, **584**: Article No. 127592

152 Zhao M M, Wang D, Li M H, Gao N, Qiao J F. A new Q-function structure for model-free adaptive optimal tracking control with asymmetric constrained inputs. *International Journal of Adaptive Control and Signal Processing*, 2024, **38**(5): 1561−1578

153 Wang T, Wang Y J, Yang X B, Yang J. Further results on optimal tracking control for nonlinear systems with nonzero equilibrium via adaptive dynamic programming. *IEEE Transactions on Neural Networks and Learning Systems*, 2023, **34**(4): 1900−1910

154 Li D D, Dong J X. Approximate optimal robust tracking control based on state error and derivative without initial admissible input. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2024, **54**(2): 1059−1069

155 Zhang H G, Luo Y H, Liu D R. Neural-network-based near-optimal control for a class of discrete-time affine nonlinear systems with control constraints. *IEEE Transactions on Neural Networks*, 2009, **20**(9): 1490−1503

156 Marvi Z, Kiumarsi B. Reinforcement learning with safety and stability guarantees during exploration for linear systems. *IEEE Open Journal of Control Systems*, 2022, **1**: 322−334

157 Zanon M, Gros S. Safe reinforcement learning using robust MPC. *IEEE Transactions on Automatic Control*, 2021, **66**(8): 3638−3652

158 Yang Y L, Vamvoudakis K G, Modares H, Yin Y X, Wunsch D C. Safe intermittent reinforcement learning with static and dynamic event generators. *IEEE Transactions on Neural Networks and Learning Systems*, 2020, **31**(12): 5441−5455

159 Yazdani N M, Moghaddam R K, Kiumarsi B, Modares H. A safety-certified policy iteration algorithm for control of constrained nonlinear systems. *IEEE Control Systems Letters*,

2020, **4**(3): 686−691

160 Yang Y L, Vamvoudakis K G, Modares H. Safe reinforcement learning for dynamical games. *International Journal of Robust and Nonlinear Control*, 2020, **30**(9): 3521−3800

161 Song R Z, Liu L, Xia L N, Lewis F L. Online optimal event-triggered H_∞ control for nonlinear systems with constrained state and input. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2023, **53**(1): 131−141

162 Fan B, Yang Q M, Tang X Y, Sun Y X. Robust ADP design for continuous-time nonlinear systems with output constraints. *IEEE Transactions on Neural Networks and Learning Systems*, 2018, **29**(6): 2127−2138

163 Liu S H, Liu L J, Yu Z. Safe reinforcement learning for affine nonlinear systems with state constraints and input saturation using control barrier functions. *Neurocomputing*, 2023, **518**: 562−576

164 Farzanegan B, Jagannathan S. Continual reinforcement learning formulation for zero-sum game-based constrained optimal tracking. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2023, **53**(12): 7744−7757

165 Marvi Z, Kiumarsi B. Safe reinforcement learning: A control barrier function optimization approach. *International Journal of Robust and Nonlinear Control*, 2021, **31**(6): 1923−1940

166 Qin C B, Qiao X P, Wang J G, Zhang D H, Hou Y D, Hu S L. Barrier-critic adaptive robust control of nonzero-sum differential games for uncertain nonlinear systems with state constraints. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2024, **54**(1): 50−63

167 Xu J H, Wang J C, Rao J, Zhong Y J, Wang H Y. Adaptive dynamic programming for optimal control of discrete-time nonlinear system with state constraints based on control barrier function. *International Journal of Robust and Nonlinear Control*, 2021, **32**(6): 3408−3424

168 Jha M S, Kiumarsi B. Off-policy safe reinforcement learning for nonlinear discrete-time systems. *Neurocomputing*, 2024, **611**: Article No. 128677

169 Zhang L Z, Xie L, Jiang Y, Li Z S, Liu X Q, Su H Y. Optimal control for constrained discrete-time nonlinear systems based on safe reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2025, **36**(1): 854−865

170 Cohen M H, Belta C. Safe exploration in model-based reinforcement learning using control barrier functions. *Automatica*, 2023, **147**: Article No. 110684

171 Liu S H, Liu L J, Yu Z. Fully cooperative games with state and input constraints using reinforcement learning based on control barrier functions. *Asian Journal of Control*, 2024, **26**(2): 888−905

172 Zhao M M, Wang D, Song S J, Qiao J F. Safe Q-learning for data-driven nonlinear optimal control with asymmetric state constraints. *IEEE/CAA Journal of Automatica Sinica*, 2024, **11**(12): 2408−2422

173 Liu D R, Li H L, Wang D. Neural-network-based zero-sum game for discrete-time nonlinear systems via iterative adaptive dynamic programming algorithm. *Neurocomputing*, 2013, **110**: 92−100

174 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

175 Zhong X N, He H B, Wang D, Ni Z. Model-free adaptive control for unknown nonlinear zero-sum differential game. *IEEE Transactions on Cybernetics*, 2018, **48**(5): 1633−1646

176 Wang Y, Wang D, Zhao M M, Liu N, Qiao J F. Neural Q-learning for discrete-time nonlinear zero-sum games with adjustable convergence rate. *Neural Networks*, 2024, **175**: Article No. 106274

177 Zhang Y W, Zhao B, Liu D R, Zhang S C. Event-triggered control of discrete-time zero-sum games via deterministic policy gradient adaptive dynamic programming. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2022, **52**(8): 4823−4835

178 Lin M D, Zhao B, Liu D R. Policy gradient adaptive dynamic programming for nonlinear discrete-time zero-sum games with unknown dynamics. *Soft Computing*, 2023, **27**: 5781−5795

179 Wang Ding, Zhao Hui-Ling, Li Xin. Adaptive critic control for wastewater treatment systems based on multiobjective 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)

180 Yang Q M, Cao W W, Meng W C, Si J. Reinforcement-learning-based tracking control of waste water treatment process under realistic system conditions and control performance requirements. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2022, **52**(8): 5284−5294

181 Yang R Y, Wang D, Qiao J F. Policy gradient adaptive critic design with dynamic prioritized experience replay for wastewater treatment process control. *IEEE Transactions on Industrial Informatics*, 2022, **18**(5): 3150−3158

182 Qiao J F, Yang R Y, Wang D. Offline data-driven adaptive critic design with variational inference for wastewater treatment process control. *IEEE Transactions on Automation Science and Engineering*, 2024, **21**(4): 4987−4998

183 Sun B, Kampen E J V. Incremental model-based global dual heuristic programming with explicit analytical calculations applied to flight control. *Engineering Applications of Artificial Intelligence*, 2020, **89**: Article No. 103425

184 Zhou Y, Kampen E J V, Chu Q P. Incremental model based online heuristic dynamic programming for nonlinear adaptive tracking control with partial observability. *Aerospace Science and Technology*, 2020, **105**: Article No. 106013

185 Zhao Zhen-Gen, Cheng Lei. Performance optimization for tracking control of fixed-wing UAV with incremental Q-learning. *Control and Decision*, 2024, **39**(2): 391−400
(赵振根, 程磊. 基于增量式 Q 学习的固定翼无人机跟踪控制性能优化. 控制与决策, 2024, **39**(2): 391−400)

186 Cao W W, Yang Q M, Meng W C, Xie S Z. Data-based robust adaptive dynamic programming for balancing control performance and energy consumption in wastewater treatment process. *IEEE Transactions on Industrial Informatics*, 2024, **20**(4): 6622−6630

187 Fu Y, Hong C W, Fu J, Chai T Y. Approximate optimal tracking control of nondifferentiable signals for a class of continuous-time nonlinear systems. *IEEE Transactions on Cybernetics*, 2022, **52**(6): 4441−4450

王 鼎 北京工业大学信息科学技术学院教授. 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 School of Information Science and Technology, Beijing University of Technology. His research interest covers reinforcement learning and intelligent control. Corresponding author of this paper.)

赵明明 北京工业大学信息科学技术学院博士研究生. 主要研究方向为强化学习和智能控制.
E-mail: zhaomm@emails.bjut.edu.cn
(**ZHAO Ming-Ming** Ph.D. candidate at the School of Information Science and Technology, Beijing University of Technology. His research interest covers reinforcement learning and intelligent control.)

刘德荣 南方科技大学自动化与智能制造学院教授. 主要研究方向为强化学习和智能控制.
E-mail: liudr@sustech.edu.cn
(**LIU De-Rong** Professor at the School of Automation and Intelligent Manufacturing, Southern University of Science and Technology. His research interest covers reinforcement learning and intelligent control.)

乔俊飞 北京工业大学信息科学技术学院教授. 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 School of Information Science and Technology, Beijing University of Technology. His research interest covers intelligent control of wastewater treatment processes, structure design and optimization of neural networks.)

宋世杰 西南交通大学智慧城市与交通学院讲师. 主要研究方向为强化学习和智能控制.
E-mail: shijie.song@swjtu.edu.cn
(**SONG Shi-Jie** Lecturer at the Institute of Smart City and Intelligent Transportation, Southwest Jiaotong University. His research interest covers reinforcement learning and intelligent control.)