国防科技大学学报
Journal of National University of Defense Technology
ISSN 1001-2486,CN 43-1067/T

《国防科技大学学报》网络首发论文

题目：模型未知系统在线强化学习控制：理论、方法及挑战
作者：张皓然，赵春晖，吴争光
收稿日期：2025-06-30
网络首发日期：2025-11-06
引用格式：张皓然，赵春晖，吴争光．模型未知系统在线强化学习控制：理论、方法及挑战[J/OL]．国防科技大学学报．
https://link.cnki.net/urlid/43.1067.t.20251106.1440.002

国防科技大学学报
JOURNAL OF NATIONAL UNIVERSITY OF DEFENSE TECHNOLOGY

doi: 10.11887/j.issn.1001-2486.25060038

模型未知系统在线强化学习控制：理论、方法及挑战

张皓然，赵春晖*，吴争光

（浙江大学 控制科学与工程学院 工业控制技术全国重点实验室， 浙江 杭州 310027）

摘　要：在智能制造、航空航天、机器人等领域，系统动态模型未知的问题普遍存在，严重制约了传统基于模型控制方法的应用。强化学习作为一种数据驱动控制方法，具备通过与环境交互实现控制策略学习优化的能力，在应对模型未知场景下的最优控制任务中展现出广阔前景。围绕连续时间系统中的动态模型未知问题，通过结合工业实例、理论分析结果等方式，回顾了通用强化学习算法发展脉络及在模型已知场景的应用，梳理了基于模型强化学习、离策略积分强化学习和 Q 学习等模型未知场景的代表性方法，介绍了基于 Lyapunov 的理论分析工具及相关假设，重点讨论了信息不完备场景下的强化学习决策大模型、安全强化学习以及稳定性与鲁棒性增强等前沿方向及现有方法面临的挑战。

关键词：强化学习；数据驱动控制；模型未知系统；在线强化学习；智能控制

中图分类号：TP13；TP181　　**文献标志码：**A

A review of online reinforcement learning control for systems with unknown models: theory, methods, and challenges

ZHANG Haoran, ZHAO Chunhui, WU Zhengguang*

(State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China)

Abstract: In the fields of intelligent manufacturing, aerospace, and robotics, control systems often operate under unknown dynamics. This significantly limits the effectiveness of traditional model-based control methods. RL (Reinforcement learning), as a data-driven intelligent control approach, enables policy learning and optimization through interaction with the environment, showing great potential for solving optimal control problems in such model-unknown scenarios. This survey focuses on the issue of unknown dynamic models in continuous-time systems and reviews the development of general reinforcement learning algorithms and their application in model-known scenarios through industrial examples and theoretical analysis methods. It also summarizes representative methods for model-unknown scenarios, such as model-based RL, off-policy integral RL, and Q-learning approaches. The survey introduces Lyapunov-based theoretical analysis tools and important assumptions. It discusses cutting-edge topics such as RL under partial observability using large language models, safe RL, and stability and robustness enhanced RL, while highlighting the challenges faced by existing methods.

Keywords: reinforcement learning; data-driven control; model-unknown systems; online reinforcement learning; intelligent control

收稿日期：2025-06-30

基金项目：浙江省“尖兵”“领雁”研发攻关计划基金资助项目（2024C01163）；国家自然科学基金资助项目（62133003）；工业控制技术全国重点实验室浙大专项资助项目（ICT2025C01）；工业控制技术全国重点实验室开放课题资助项目（ICT2025B07）

第一作者：张皓然（1996—），男，江苏徐州人，助理研究员，博士，E-mail：0625495@zju.edu.cn

*__通信作者：__赵春晖（1979—），女，山东莱州人，教授，博士，博士生导师，E-mail：chhzhao@zju.edu.cn

引用格式：张皓然，赵春晖，吴争光. 模型未知系统在线强化学习控制：理论、方法及挑战[J]. 国防科技大学学报.

Citation: ZHANG H R, ZHAO C H, WU Z G. A review of online reinforcement learning control for systems with unknown models: theory, methods, and challenges[J]. Journal of National University of Defense Technology.

当前，全球正处于新一轮工业革命和数字化转型的浪潮中，我国相继出台了《中国制造 2025》《“十四五”智能制造发展规划》等战略性文件，明确要求依托新一代信息技术推动传统产业向高端化、智能化和绿色化转型[1-2]。这些国家战略和政策为工业自动化、智能制造及高性能控制系统的发展提供了坚实的保障，同时也为提升国家竞争力和产业升级指明了方向。随着人工智能、大数据、云计算和物联网等前沿技术的不断突破，未来工业系统将呈现出高度数字化和智能化的特征，这对控制系统提出了前所未有的挑战：如何在充满动态不确定、系统模型未知的复杂环境中，实现稳定、鲁棒、安全、实时且高性能的控制，已经成为亟待解决的重要课题[3-5]。

回顾控制理论的发展历程，自 20 世纪 40 年代美国数学家诺伯特·维纳系统性地提出控制科学思想以来，该领域不断推陈出新，逐步形成了涵盖线性、非线性、最优、鲁棒、自适应以及智能控制等多种理论体系[6]。图 1 展示了控制科学的简要发展历程[7]。其中，诞生于冷战时期的最优控制理论不仅为航空航天等关键领域提供了强有力的技术支撑，也推动了现代控制技术的快速发展。例如，在阿波罗登月计划中，美国采用最优控制方法成功实现了姿态与轨道的高精度调控，确保了人类首次登月任务的成功[8]。在理论层面，最优控制可在精确建模前提下实现性能指标的最优化。然而，在实际工程中，无论是大型流程工业还是复杂装备系统，模型不确定性、环境扰动和突发故障等因素普遍存在，使得传统依赖精确建模的方法常常面临失效的风险。

图 1　控制科学的发展历程

Fig.1 The development of control science

历史上的多个案例，如 1967 年美国宇航局 X-15-3 高性能飞行器事故[9]以及 1987 年瑞典萨博公司的鹰狮（JAS39）战斗机原型事故[10]，都是由于控制系统未能充分应对系统未知动态和非线性效应，导致飞行器坠毁甚至飞行员伤亡。这些历史案例揭示了传统控制方法在应对未知动态和复杂非线性时的脆弱性，凸显出对新型控制技术的迫切需求。

在这种情况下，基于数据驱动的智能控制方法应运而生。近年来，随着机器学习和人工智能技术的迅速发展，强化学习（reinforcement learning）作为一种基于交互数据进行试错优化的智能控制方法，正逐渐成为解决复杂最优控制问题的重要工具。强化学习目前已在多个领域已取得突破性成果（图 2），例如：2016 年 AlphaGo 在围棋比赛中战胜世界冠军李世石[11]，AlphaFold 在蛋白质结构预测中取得革命性进展并助力获得 2024 年诺贝尔化学奖[12]，以及 ChatGPT 和 DeepSeek 等大语言模型通过强化学习的训练显著提升了模型的智能水平[13]。这些成功案例不仅引发了广泛关注，也极大地推动了强化学习在工程控制中的应用探索。与传统方法相比，强化学习不依赖系统显式系统数学模型，具备良好的自适应性和可扩展性，在面对动态模型未知的非线性系统和复杂环境时展现出显著优势。

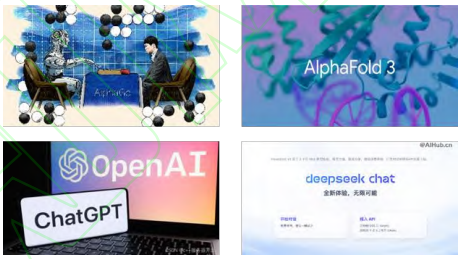

图 2　强化学习的成功应用

Fig.2 Successful applications of reinforcement learning

尽管强化学习在上述领域取得了突破性进展，其在实际工程控制中的应用落地仍然面临诸多挑战。一方面，现有强化学习方法普遍依赖高保真仿真环境与大量训练时间，在真实工业系统中往往难以复现；另一方面，强化学习在策略收敛性、稳定性和安全性方面尚缺乏理论保障，智能体的行为缺乏可解性。以上两点原因导致现有方法在训练时必须重置环境、迭代训练，只能归类于离线方法。因此，这里主要探讨如何将强化学习与控制理论相结合，构建既能通过数据驱动实现在线学习更新，又能在实时环境中确保闭环系统稳定性的最优控制，即在线强化学习方法。

近年来，已有若干国内外综述文章对强化学习在控制系统中的应用进行了总结。例如，文献[14]回顾了以 Q 学习为代表的强化学习算法在数据驱动控制中的应用，文献[15]探讨了自适应动态规划方法在离散时间系统以及在信息物理系统上的应用，文献[16]聚焦于离线学习场景下的深度强化学习发展趋势，文献[17]则讨论了强化学习与大语言模型相结合的最新进展。这些工作为理解强

化学习的基本原理及其工程实践提供了有益参考。然而，通过对上述文献的梳理可以发现，现有综述仍在以下角度存在空缺：1）大多数研究集中于离散时间系统或随机系统，而针对在工业控制中更具代表性的连续时间非线性系统相关综述仍较为稀缺；2）尚缺乏从控制理论视角出发，系统分析强化学习在实际部署中面临的稳定性、安全性与实时性问题；3）大部分综述以算法发展为主线，对模型未知场景带来的问题与挑战缺少详细深入的解释。

鉴于上述背景，本文围绕连续时间系统中动态模型未知场景下的在线强化学习方法展开系统综述，深入浅出地介绍了控制系统中模型未知问题的工程背景与系统特性、通用强化学习算法、面向模型未知场景的在线强化算法、理论分析与算法性能边界，以及当前方法在工程落地中的难点与发展趋势。图 3 展示了主要综述的关键方法和挑战，主要包含基于模型强化学习、离策略积分强化学习以及 Q 学习三种代表性算法，同时分析了这些方法在信息不完备、安全保障机制与稳定性等 方面的共性问题，最后讨论并展望了解决这些问题的可行方法及未来发展趋势。

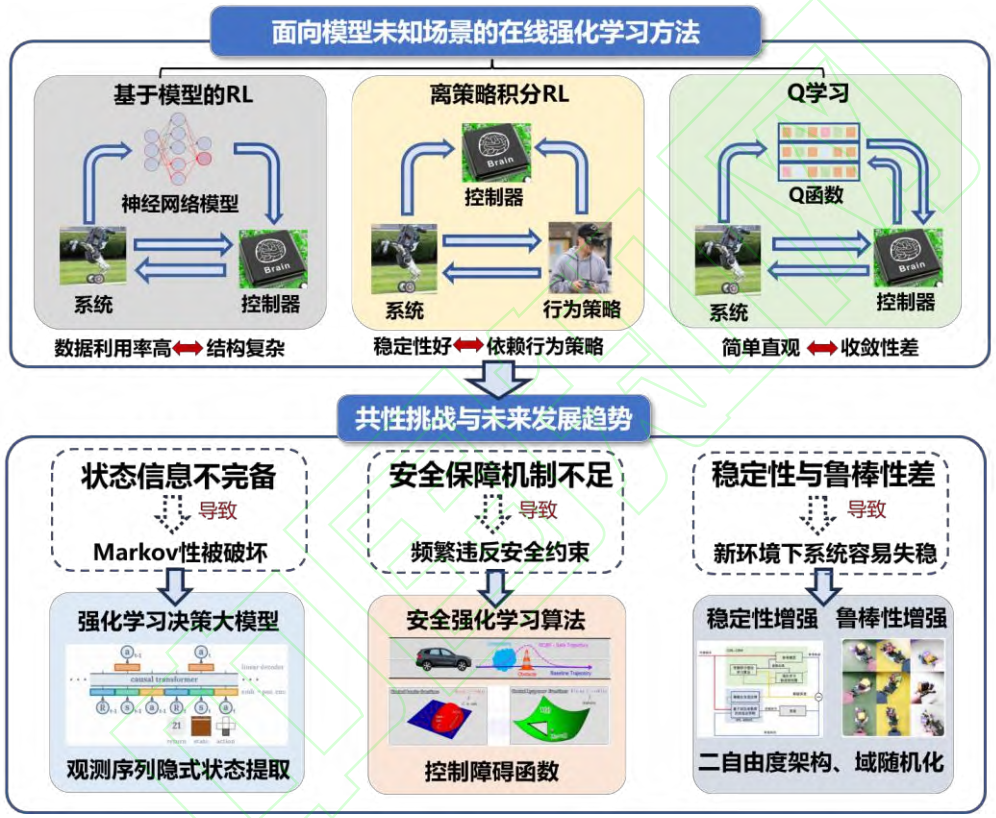

图 3　面向模型未知场景的在线强化学习控制

Fig.3　Existing methods of online reinforcement learning for model-unknown systems

1　控制系统中的模型未知问题：挑战与传统方法

在复杂系统的控制过程中，系统动态模型的准确性直接关系到控制性能和闭环稳定性。然而，实际工程中常常面临模型结构未知、参数时变、近似误差等挑战，导致依赖精确模型设计的传统控制方法在复杂环境下难以有效应用。此类“模型未知”的动态系统广泛存在于流程工业、航空航天、机器人等领域[18-19]。为应对这类系统，控制理论发展出鲁棒控制、自适应控制、数据驱动控制等方法，以期在模型未知情况下仍能保障系统的性能与安全性。本节主要从几个典型案例出发，沿着线性系统、非线性仿射系统以及高阶非线性系统的脉络，梳理并分析控制系统中的模型未知问题，并简要介绍一些对应的传统方法。

1.1　典型线性系统：以 F16 战斗机攻角调节系统为例

F-16 战斗机是一种高机动飞行器，其飞行控制系统需在不同工况下完成姿态调节、航向控制等复杂任务。其中，攻角（angle of attack）调节是关键子系统之一，其动力学在一定飞行包线内可近似为如下线性系统模型[20-21]，

$$\dot{\boldsymbol{x}} = A\boldsymbol{x} + Bu, \tag{1}$$

其中，系统状态 $\boldsymbol{x} = [x_1, x_2, x_3]^\top$ 分别包含攻角、俯仰率和升降舵偏转角，控制 u 代表升降舵控制输入的电压信号。当参数已知时，系统矩阵为

$$\begin{cases} A = \begin{bmatrix} -1.01887 & 0.90506 & -0.00215 \\ 0.82225 & -1.07741 & -0.17555 \\ 0 & 0 & -1 \end{bmatrix} \\ B = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}. \end{cases} \tag{2}$$

针对该系统，线性二次调节与极点配置等方法在模型已知时可以提供良好的控制性能[22-23]。然而，实际飞行环境中，气动系数随飞行速度、高度和姿态的变化显著，舵面存在迟滞与饱和，且受到强烈风扰与执行器动态影响，导致式（1）中的系统矩阵 (A, B) 难以准确获知。此外，这些非线性和耦合动态也很难统一纳入模型中建模。

在这种情况下，控制领域提出了包括 μ－合成、H 无穷控制等多种鲁棒控制方法[24]。这类方法的基本思想是将式（1）系统分解成如下完全已知的标称系统和未知动态的形式：

$$\dot{\boldsymbol{x}} = (A + \Delta A)\boldsymbol{x} + (B + \Delta B)\boldsymbol{u} + f(\boldsymbol{x}, \boldsymbol{u}), \tag{3}$$

其中，ΔA 和 ΔB 代表系统模型线性未知部分，f 则代表未知的非线性动态。对于鲁棒控制而言，只要未知的部分是有界的，那么总可以找到一个能够容忍所有不确定性的固定增益控制器。这类方法虽然结构简单有效，但控制性能较为保守。

另一类方法是自适应控制，如模型参考自适应控制与自校正控制[25]。该类方法通过在线调节控制增益以适应系统参数变化，分为间接（先辨识后控制）与直接（基于误差调节增益）两类。尽管自适应控制不依赖完整模型信息，但其暂态性能波动较大，容易引起振荡[26]。

1.2 典型非线性仿射系统：以连续搅拌反应釜为例

连续搅拌反应釜（continuous stirred tank reactor, CSTR）是过程工业中的关键设备，广泛应用于化工、制药和材料合成等行业。CSTR 系统的动力学表现出强非线性、强耦合及多扰动特征。以一阶放热反应为例，在忽略复杂流体力学因素下，系统可建模为如下二阶非线性系统[20]：

$$\begin{aligned} \dot{\boldsymbol{x}} &= f(\boldsymbol{x}) + g(\boldsymbol{x})u + \boldsymbol{d} \\ &= \begin{bmatrix} -x_1 + D(1 - x_1)\mathrm{e}^{\gamma x_2 (\gamma + x_2)^{-1}} \\ -x_2 - bD(1 - x_1)\mathrm{e}^{\gamma x_2 (\gamma + x_2)^{-1}} - \beta x_2 \end{bmatrix} \\ &\quad + \begin{bmatrix} 0 \\ \beta \end{bmatrix} u + \boldsymbol{d}, \end{aligned} \tag{4}$$

其中，系统状态 $\boldsymbol{x} = [x_1, x_2]^\top$ 分别代表反应釜中反应物的浓度和温度，控制变量 u 代表冷却水温度，$\boldsymbol{d}$ 代表未知扰动，D 是表示反应速率与流出速率比值的达姆科勒数，b 代表绝热温升，γ 代表无量纲形式的活化能，β 为传热系数，是热传递速率与进料热容流量之比。

CSTR 在实际运行中常常面临建模误差与参数漂移问题，使得动态函数 $f(x)$ 和 $g(x)$ 中的参数难以提前获取，更不用说式（4）系统中还有包含外部扰动和未建模动态。在这种情况下，传统反馈线性化和自适应控制在扰动与非结构不确定性下的性能往往难以保障。目前对于这类系统，模糊控制、神经网络控制等智能控制方法成为传统方法的有效补充[27]。这类方法基于通用逼近能力（如径向基函数网络、模糊规则系统），可以在无需精确模型的条件下近似系统动态，实现模型未知情况下的有效控制。

1.3 典型高阶非线性系统：以多自由度机械臂系统为例

机械臂作为典型的多入多出非线性系统，广泛应用于现代工业制造、仓储物流及服务机器人等领域。由于机械臂关节摩擦、柔性连接、负载变化等因素的影响，其动力学模型往往难以准确获取，导致传统模型驱动控制方法在实际应用中面临严重挑战。以典型的 n 自由度刚性关节机械臂为例，其动力学模型可表示为如下欧拉-拉格朗日系统的形式[28-29]：

$$M(\boldsymbol{q})\ddot{\boldsymbol{q}} + V_m(\boldsymbol{q}, \dot{\boldsymbol{q}})\dot{\boldsymbol{q}} + F(\dot{\boldsymbol{q}}) + g(\boldsymbol{q}) + \boldsymbol{\tau}_d = \boldsymbol{\tau}, \tag{5}$$

其中，$\dot{\boldsymbol{q}} \in \mathrm{R}^n$ 和 $\boldsymbol{q} \in \mathrm{R}^n$ 分别代表关节角度和速度，$\boldsymbol{\tau} \in \mathrm{R}^n$ 是关节力矩，$M(\cdot) \in \mathrm{R}^{n \times n}$ 是惯性矩阵，$V_m(\cdot) \in \mathrm{R}^{n \times n}$ 是科里奥利与离心力矩阵，$F(\cdot) \in \mathrm{R}^n$ 代表摩擦力，$g(\cdot) \in \mathrm{R}^n$ 是重力，$\boldsymbol{\tau}_d$ 是外部扰动。

虽然上述模型理论上可以通过拉格朗日方法从机械结构推导得到，但实际中由于负载变化、摩擦非线性、柔性结构、材料老化等因素存在，式（5）系统中的 M、V、F、g 以及 τ_d 等动态存在高度不确定性。此时，基于解析模型设计的控

制器在部署后通常会发生性能下降，甚至无法满足稳定性和精度要求。

因此，除了前面提到的一些方法外，近年来大量研究聚焦于基于数据驱动或强化学习的机械臂控制方案，从而在系统动力学模型未知的情况下实现高性能控制。例如，无模型自适应控制可以将机械臂系统（5）转化为如下动态线性化数据模型[30]：

$$\Delta \boldsymbol{y}(k+1)=\boldsymbol{\Phi}_{\boldsymbol{c}}(k)\Delta \boldsymbol{u}(k), \tag{6}$$

其中，k 代表离散化的时间步，$\boldsymbol{\Phi}_{\boldsymbol{c}}(k)$ 代表伪雅可比矩阵，输出变化定义为 $\Delta \boldsymbol{y}(k+1)=\boldsymbol{q}(k+1)-\boldsymbol{q}(k)$，输入变化定义为 $\Delta \boldsymbol{u}(k+1)=\boldsymbol{\tau}(k+1)-\boldsymbol{\tau}(k)$。实际使用时只需要设计算法对 $\boldsymbol{\Phi}_{\boldsymbol{c}}(k)$ 进行实时估计，再将估计值代入控制律中即可完成机械臂的实时控制。相比于传统方法，该方法无需依赖显式模型（5），具备较强的通用性与实现简便性。

综上所述，从线性系统、典型非线性系统到高阶复杂系统，模型未知问题广泛存在并严重制约传统控制方法的性能与适应能力。尽管鲁棒控制、自适应控制等方法在一定程度上可提升系统对模型误差的容忍性，但其仍依赖部分模型知识，在实际复杂环境下往往难以高效、快速地完成控制任务。随着人工智能方法与算力的发展，强化学习作为一种无需依赖精确模型、具备交互式学习能力的控制方法，正在成为解决模型未知问题的重要路径。下一节中将从基础概念出发，系统介绍强化学习的理论框架与通用算法，并重点分析其与传统最优控制之间的联系，为后续面向最优控制的技术演进奠定基础。

2 从交互式学习到闭环控制：通用强化学习方法

上一节用举例的方式介绍了控制系统中的动态模型未知问题。针对这一难题，现有方法虽然积累了丰富成果，但在适应系统复杂性和多样性方面仍存在明显局限。近年来，人工智能技术的兴起为控制系统带来了新的设计思路，尤其是强化学习凭借无需依赖系统模型、数据驱动交互式学习等特性，正快速成为应对未知动态系统最优控制问题的潜在候选方案。强化学习的兴起不仅来源于人工智能领域的算法演进，更得益于其在通用决策任务中展现出的出色表现。为了更好地理解这一技术的基本原理与研究基础，本节将围绕强化学习的核心理论展开，介绍通用的强化学习算法及其发展脉络。

强化学习是一种通过智能体（agent）与环境（environment）不断交互、试错优化决策策略，最终实现累计奖励最大化的机器学习方法[31]。其基本理论通常由马尔可夫决策过程（Markov decision process, MDP）来形式化描述。MDP 由状态空间 S，动作空间 A，状态转移概率 $P(s'|s,a)$ 以及奖励函数 $r(s,a)$ 构成。在 MDP 框架下，智能体在每个时刻 t 根据当前状态 s_t 和策略 $\pi(a_t|s_t)$ 选择动作 a_t，环境则根据 $P(s_{t+1}|s_t,a_t)$ 转移到新的状态 s_{t+1} 并反馈奖励 $r_t=r(s_t,a_t)$。这一过程会随着时间的累积而不断进行，而智能体的目标则是不断优化策略函数 $\pi(\cdot|\cdot):S\to A$，使得如下累计折扣奖励最大化[32]：

$$J=\mathrm{E}_{s_t\sim P,a_t\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right], \tag{7}$$

其中 $\gamma\in(0,1]$ 为折扣因子，用来平衡未来奖励对 J 的影响[31]。

从发展脉络来梳理强化学习算法，其大体可以分为传统强化学习和深度强化学习。前者使用线性或浅层模型作为函数逼近器，主要用于离散或低维的状态－动作空间问题，后者则利用深度神经网络处理高维连续的状态－动作空间问题。依据算法优化目标的不同强化学习也可以分为值函数法、策略梯度法以及联合两者的 Actor-Critic 法[7,33]。图 4 简要展示了强化学习方法的发展脉络，可以看到上述几类方法在历史上交替演进，并逐步推动了强化学习理论的成熟和应用。

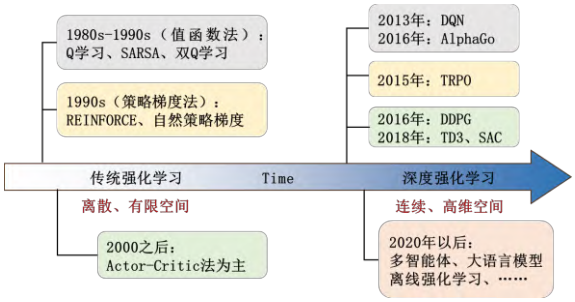

图 4 通用强化学习算法发展脉络

Fig.4 The development of general RL algorithms

2.1 传统强化学习

最早期的强化学习方法主要面向离散空间的场景，例如象棋和多臂老虎机等[31]。在这种情况下，S 和 A 都只包含有限个元素，因此可以用一张“表格”来表达所有可能的状态－动作对 (s,a) 并存储 (s,a) 对应的值函数 $Q(s,a)$。其中，值函数用于评估策略的性能，当计算得到所有 (s,a) 对

应的值函数时，就可以通过查表得到最优策略。20 世纪 80 年代，Watkins 博士提出的 Q 学习算法是该领域的里程碑[34]。作为一个经典的值函数法，其更新公式如下，

$$Q(s,a)\leftarrow Q(s,a)+\alpha\big[r(s,a)+\gamma\max_{a'}Q(s',a')-Q(s,a)\big], \tag{8}$$

其中 $\alpha>0$ 为学习率。这种方法通过对状态－动作值函数 $Q(s,a)$ 的迭代更新，使智能体能够在有限动作空间内学习最优策略。随后，Sarsa 算法[31]、双 Q 学习（Double Q-learning）[35]等方法相继出现，针对 Q 学习的过估计等问题进行了改进。然而，Q 学习及其变种主要适用于离散动作问题，难以直接应用于连续动作空间的任务。

稍晚一点出现的是策略梯度法，这类方法直接对策略进行参数化，既可以用在离散动作空间，也可以用于连续动作空间。20 世纪 90 年代，REINFORCE 算法[36]、策略梯度法[37]，以及自然策略梯度法[38]被相继提出。以 Williams 提出的 REINFORCE 为例，其首先构造参数化策略 $\pi_\theta(a|s)$，利用蒙特卡洛方法估计累积奖励，并通过梯度上升更新策略参数 θ [36]：

$$\theta\leftarrow\theta+\alpha\mathrm{E}_{\pi_\theta}\big[\nabla\log\pi_\theta(a|s)G\big], \tag{9}$$

其中 G 表示从当前时刻起的累计奖励。

尽管策略梯度方法具有直接优化策略的优势，但其存在方差高、效率低等问题，因此后来又出现了将值函数法与策略梯度相结合的 Actor-Critic 法。该框架中，Critic 部分负责评估策略的性能，而 Actor 部分则依据 Critic 的反馈更新策略，其梯度更新表达式通常写成如下形式[7,33]，

$$\theta\leftarrow\theta+\alpha\mathrm{E}_{\pi_\theta}\big[\nabla\log\pi_\theta(a|s)A(s,a)\big], \tag{10}$$

其中 $A(s,a)$ 代表由 Critic 估计的优势函数，反映了某个动作相对于平均水平的优劣。相比于公式（9），该方法能实现单步更新，因此提高了学习效率。Actor-Critic 已经基本成为先进强化学习算法的标准框架。

2.2 深度强化学习

随着深度学习技术的发展，深度强化学习利用深度神经网络作为函数逼近器，成功克服了传统方法在处理高维状态和动作空间时的困难[39]。2013 年，Mnih 等人提出了深度 Q 网络（deep Q network, DQN）算法，将卷积神经网络与 Q 学习相结合，并引入经验回放机制和固定目标网络来稳定训练，使得计算机在大多数雅达利游戏中超过了人类专家的水平[40]。此后，DDQN[41]、Dueling-DQN[42]、Rainbow[43]等一系列变体不断涌现，进一步提升了算法性能。2016 年，AlphaGo 横空出世，其算法基础就是 DQN 和蒙特卡洛树搜索[11]。

DQN 及其改进算法仍属于值函数法，因此更擅长处理游戏棋牌这类离散动作空间问题。在实际工程控制中，许多任务（如机器人控制、自动驾驶等）同时涉及连续状态空间和动作空间，这类问题仍需采用策略梯度或 Actor-Critic 方法。2015 年，Schulman 等人提出了信赖域策略优化（trust region policy optimization, TRPO）算法[44]。该算法属于策略梯度法，其通过约束策略更新的步长提高深度神经网络策略训练的稳定性。此外，Silver 等人提出了确定性策略梯度法（deterministic policy gradient, DPG）[45]，该算法直接对确定性策略 $\mu_\theta(s)$ 进行优化，其更新方式为

$$\theta\leftarrow\theta+\alpha\mathrm{E}_s[\nabla_\theta\mu(s)\nabla_a Q(s,a)|_{a=\mu(s)}]. \tag{11}$$

相比于随机策略的方法（10），DPG 在无需连续动作空间中对动作进行采样，因此通常更加高效。2016 年，Lillicrap 等人在（11）基础上提出了深度确定性策略梯度法（deep deterministic policy gradient, DDPG）。作为一种结合了 DQN、DPG 以及经验回放机制的 Actor-Critic 算法，DDPG 在许多连续控制任务中表现优异，但仍存在过估计和训练不稳定问题[32]。为了解决这些不足，学界又在 2018 年分别提出了双延迟深度确定性策略梯度（twin delayed deep deterministic policy gradient, TD3）[46]和柔性 Actor-Critic (soft actor-critic, SAC)[47]。其中，TD3 通过引入延迟更新策略和双 Q 网络，有效缓解了 Q 学习算法常见的过估计问题。而 SAC 则在策略目标中引入最大熵正则项，不仅鼓励策略探索，还能提高样本效率和训练稳定性。这两种算法在处理高维连续动作控制任务时表现出色，是该领域的 SOTA 方法。

近几年，深度强化学习又分别在多智能体系统[48]、离线强化学习[16]以及训练大语言模型[49]等场景取得了一定的成果。然而，实际工业场景（如自动驾驶、机器人、工业自动化和智能制造等）对实时性、稳定性和安全性的要求远高于前面提到的娱乐和商业应用环境。在这些实际工程领域的场景中，当前强化学习算法所取得的成果往往难以直接移植，原因在于真实环境中难以构建高保真仿真平台，也无法保证长时间迭代训练的稳定性与安全性。因此，过去二十年间，也有

许多学者致力于发展能够解决最优控制问题的强化学习算法。

3 面向模型已知场景最优控制的强化学习方法

最优控制旨在通过设计控制律使动态系统的性能指标达到全局最优，其数学本质可归结为在约束条件下求解泛函极值问题。20 世纪 50 年代，苏联学者庞特里亚金和美国数学家贝尔曼分别提出了极小值原理和动态规划，奠定了现代最优控制的理论基础[22]。虽然动态规划基于贝尔曼最优性原理构建了严密的数学框架，但其逆向计算模式在状态空间维度升高时面临“维数灾难”问题，且依赖精确模型的离线计算难以实时应用。强化学习通过数据驱动的交互式学习为这一难题提供了新思路。1974 年，Werbos 结合传统动态规划与强化学习思想，提出了自适应动态规划（Adaptive Dynamic Programming, ADP）的基本框架[50-51]。其核心在于利用神经网络等函数近似结构在线逼近值函数与控制律，结合动态规划的优化思想与强化学习的自适应能力，从而按时间正向求解最优控制问题。

最优控制与强化学习虽分属控制理论与机器学习领域，但其核心目标均是通过优化性能指标实现动态系统的最优控制或决策。两者的底层逻辑是相通的：最优控制通过最小化代价函数设计控制器，强化学习则通过最大化累积奖励优化智能体策略，动态规划和贝尔曼方程为两者提供了统一的优化框架[52-54]。表 1 进一步总结对比了两者的差异与关联。同时，下图 5 给出了两者在术语方面的对照关系。

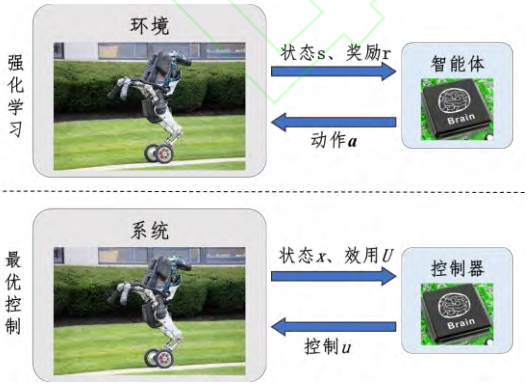

图 5　强化学习与最优控制术语对照

Fig.5 Terminology comparison of RL and optimal control

表 1　强化学习与最优控制之间的关联

Tab.1　Relationship between RL and optimal control

特性	强化学习	最优控制
优化目标	最大化累积奖励	最小化代价函数
系统建模	基于 MDP，允许模型未知	基于动态系统，要求模型已知
求解方式	基于随机逼近理论，需在线迭代学习	依赖解析方法，需离线计算
稳定性	通过仿真、实验验证	有严格数学证明
术语对应	环境、智能体、奖励函数	系统、控制器、效用函数
应用场景	金融、游戏等模型复杂或未知的探索性场景	工业控制、航空航天等模型已知的高可靠性场景

近几十年来，基于强化学习的最优控制理论发展经历了从离散时间系统到连续时间系统、从最优镇定到最优跟踪的范式演进[52-54]。与前文提到的传统强化学习与深度强化学习相比，ADP 更注重控制问题的实时性、稳定性与安全性，尤其在处理各类动态系统的调节或跟踪控制时展现出独特优势，接下来将分别梳理相关工作。

3.1 基于强化学习的最优调节

动态规划的数学基础源于贝尔曼最优性原理。对于离散时间动态系统，贝尔曼方程通过逆向递推求解最优值函数[22]：

$$V^*(x(k))=\min_{u(k)}\{U(x(k),u(k))+\gamma V^*(x(k+1))\},\tag{12}$$

其中，$x(k)$ 代表系统在 k 时刻的状态，$u(k)$ 代表控制器在 k 时刻计算的控制信号，U 是正定的效用函数（通常为 $x^\top Sx+u^\top Ru$ 的形式），V^* 代表最优值函数，$x(k+1)=F(x(k),u(k))$ 代表系统动态。和强化学习最大化累计奖励相反，控制场景中通常习惯把问题描述为最小化的形式，两者其实在数学上是等价的[55]。

求解上述方程的计算复杂度随状态维度呈指数增长，且依赖离线逆向计算，难以实时应用。为此，Werbos 提出通过函数逼近器分别得到值函数和控制律的近似：$\hat{V}(x)$ 与 $\hat{u}(x)$，并在线更新权重以最小化贝尔曼残差[50-51]：

$$E=U(x,\hat{u})+\gamma\hat{V}(F(x,\hat{u}))-\hat{V}(x),\tag{13}$$

这一方法将方程（12）的逆向递推转化为正向迭代，奠定了基于 ADP 的数据驱动最优控制基础。具体地，Werbos 提出了启发式动态规划[56]和二次启发式动态规划[57]，分别通过神经网络逼近值函

数及其梯度。后续又发展了类似 Q 学习的控制依赖启发式动态规划，通过将控制信号也纳入 Critic 网络以增强适应性[58]。进一步地，Prokhorov 等人提出全局二次启发式规划及其变体[59]，通过同时逼近值函数及其梯度以提升精度。以上方法都可以归类为 Actor-Critic 结构，为降低复杂度，Padhi 等人提出单网络自适应 Critic 方法[60]，通过单个 Critic 网络联合优化值函数与控制律，简化了计算架构。

而在连续时间动态系统中，最优控制需通过哈密尔顿 - 雅可比 - 贝尔曼（Hamilton–Jacobi–Bellman，HJB）方程描述[22]：

$$\min_{u\in\mathcal{A}}\left\{U(x,u)+(\nabla V^*)^{\top}F(x,u)\right\}=0. \quad (14)$$

和前面离散时间相关方法不同，针对连续时间系统最优控制发展的强化学习算法更强调在线学习和稳定性[50,61]。早在上世纪末，已有学者提出使用广义策略迭代离线近似求解上述 HJB 方程（14）[62]。2002 年，美国加州理工学院的 Murray 等人针对上述连续系统的最优控制问题提出了一种新型的迭代 ADP 方法，并在初始稳定控制的条件下证明了系统稳定性和算法收敛性[55]。2005 年，Lewis 等人采用非二次泛函的效用函数并设计离线策略迭代算法，顺利求解了控制受限的连续时间系统最优控制问题[63]。

2010 年左右，Vamvoudakis 和 Lewis 成功提出了一种针对非线性仿射系统的在线自适应 ADP 算法（synchronous policy iteration，SPI）[64]，实现了 Actor 和 Critic 网络的在线实时更新，并给出了稳定性证明。SPI 采用两个神经网络 $\hat{V}(x)=\hat{W}_c^{\top}\phi_c(x)$ 与 $\hat{u}(x)=-0.5R^{-1}g^{\top}\nabla\phi_c(x)\hat{W}_a$ 分别逼近值函数和最优控制。其中，$\hat{W}_c$ 和 $\hat{W}_a$ 分别代表 Critic 网络和 Actor 网络的权重，ϕ_c 代表基函数。

SPI 的神经网络参数更新律设计为

$$\begin{cases}\dot{\hat{W}}_c=-\alpha_1\dfrac{\sigma}{m_s^2}\left(\sigma^{\top}\hat{W}_c+U(x,\hat{u})\right),\\ \dot{\hat{W}}_a=-\alpha_2\left(F_2\hat{W}_a-\left(\dfrac{F_1\sigma^{\top}}{m_s}+\dfrac{D\hat{W}_a\sigma^{\top}}{4m_s^2}\right)\hat{W}_c\right),\end{cases} \quad (15)$$

其中，α_1 和 α_2 为学习律，m_s 为正则化项，F_1 和 F_2 为可调参数矩阵，σ 和 D 的定义可见原文[64]。尽管 SPI 不需要初始稳定控制，但一些近期的综述文章指出其参数调优较为困难[61,65-66]。2015 年左右，Liu 和 Wang 等人提出在 Critic 网络的更新律中加入辅助稳定项，在放松初始稳定控制条件的同时简化了网络结构，并保证了系统的稳定性[67-68]。目前有很多学者都采用这类改进的自适应 Critic 方法解决连续时间系统最优控制问题[69-71]。

3.2 基于强化学习的最优跟踪

相比于调节问题，跟踪控制需要系统状态 $x(t)$ 实时跟踪时变参考轨迹 $x_d(t)$，因此更加困难。由于 $x_d(t)$ 通常是非零的，因此无论如何选择控制律，代价函数总会随时间趋向于无穷大，此时控制会失去一般意义上的最优性[72-73]。所以相比于最优调节控制，基于强化学习的最优跟踪控制发展得相对较晚。2011 年左右，Zhang 等人提出了一种非线性连续时间系统自适应最优跟踪控制方法[74]，其中采用了基于循环神经网络的系统辨识器前馈控制消除跟踪误差。2014 年，Modares 等人首次采用了如下带折扣因子（$\gamma>0$）的代价函数，

$$J=\frac{1}{2}\int_t^{\infty}\mathrm{e}^{-\gamma(\tau-t)}U(x-x_d,u)d\tau, \quad (16)$$

并将参考系统的动态和被控系统的动态联合构造增广系统，从而成功解决了线性系统最优跟踪问题[75]。同年，他们又采用类似的办法解决了带有控制受限的非线性系统最优跟踪问题，并使用积分强化学习在线优化 Actor-Critic 网络权重[76]。这种通过折扣因子和增广系统将跟踪问题转化为调节问题的方法被后续很多工作所沿用，同样取得了较好的效果[77-81]。

引入折扣因子虽然解决了代价函数 J 随着时间推移变得无穷大的问题，但同样破坏了渐近跟踪性质，导致出现稳态误差[82]。近年来，也有学者积极采用其他类型的控制结构和代价函数，从而提高控制效果。例如，Na 等人利用在线辨识模型构造前馈控制消除跟踪误差[83]；Li 和 Wang 等人设计了新式的效用函数解决离散系统的跟踪问题[84-85]，通过将未来时刻的跟踪误差纳入效用函数，在保留折扣因子的同时消除了跟踪误差；Sereshki 等人在设计效用函数时显式考虑了前馈控制项，从而避免近似求解 HJB 方程[86-87]。此外，结合内模原理消除跟踪误差是目前较为流行的方法[88-92]。这类方法会在效用函数中引入类似如下的动态结构，

$$\begin{cases}J=\dfrac{1}{2}\int_t^{\infty}U(x-x_d,\Delta(s)u)d\tau,\\ \Delta(s)=s^p+a_1s^{p-1}+\cdots+a_{p-1}s+a_p,\end{cases} \quad (17)$$

其中 $\Delta(s)$ 是和参考轨迹相关的特征多项式（s 代表微分算子），一般要求 $\Delta(s)x_d=0$。可以看到

这类基于内模原理设计的方法不需要引入额外的折扣因子，也不会造成代价函数 J 随时间趋向于无穷，但要求对参考轨迹有良好的先验知识[93]。

总体来看，尽管基于强化学习的最优控制方法在理论层面取得显著进展，但大部分方法仍然要求对被控对象的动态模型有着充分的了解。从方程（12）和（14）可以看到，无论是离散还是连续时间系统，在求解过程中都假设系统动态 $F(x,u)$ 已知，这显然限制了强化学习在工程实践中的应用范围。因此，下一节将梳理面向数据驱动最优控制的在线强化学习方法，分析在系统动态模型未知情况下解决最优控制问题的最新突破。

4 面向模型未知场景最优控制的强化学习方法

正如前文所述，由于现实中大部分系统的动态模型都存在一定程度的不确定性，上一节提到的大部分通用方法在这种情况下并不适用。针对这一问题，国内外学者开展了大量工作，主要围绕如何在系统动态模型部分或完全未知的情况下，通过数据驱动的方式求解最优控制问题展开研究。

不失一般性地，以连续时间非线性仿射系统为例：

$$\dot{x} = F(x,u) = f(x) + g(x)u, \tag{18}$$

其中 $f(x)$ 和 $g(x)$ 都是未知的系统动态函数。正如在第 1 节所述，造成模型未知的原因可能是由于非线性耦合、参数不确定和未建模动态等[94]。例如，机器人关节摩擦与负载变化[95-96]、智能电网中可再生能源的随机波动[97]、化工过程中反应物浓度与催化剂活性[98-99]不确定等因素都可能造成系统动态未知。

针对形如式（18）系统的未知动态系统最优控制问题，本节将聚焦于目前三类比较流行方法：基于模型的强化学习、离策略积分强化学习以及 Q 学习。下面将按照这三大类进行归纳整理，并通过关键公式展示各自的核心思想。

4.1 基于模型的强化学习

基于模型的强化学习（model-based RL）是一种较直观的思路，即先对式（18）系统进行数据驱动建模，然后在所建立模型的基础上使用上一节提到的传统算法学习最优控制律。典型代表工作是 Bhasin 等人在 2013 年提出的 Actor–Critic-Identifier 架构[100]。相比于前文提到的 Actor–Critic 方法，该方法多了一个基于循环神经网络的系统辨识器/系统模型，其具体形式为：

$$\dot{\hat{x}} = \hat{W}_f^{\top}\sigma(\hat{V}_f^{T}\hat{x}) + g(x)u + \mu \tag{19}$$

其中，$\hat{x}$ 为式（19）辨识器的状态，$\hat{W}_f$ 和 $\hat{V}_f$ 都是神经网络参数，μ 是处理近似误差和扰动的鲁棒项。

该方法通过在线更新网络参数，使得式（19）辨识器可以逐渐逼近式（18）系统的真实动态。因此，（19）可以代替原本未知的（18）辅助 Actor 和 Critic 网络学习。注意到式（19）辨识器中假设 $g(x)$ 是已知的，实际上有很多其他辨识器设计可以规避这一点。例如，Modares 等人使用两个神经网络分别建模 $f(x)$ 和 $g(x)$，并将系统动态转换成滤波回归的形式进行学习[101]，其他类似的方法还可见文献[83,102-104]等。

总体而言，基于模型的强化学习算法基本沿用了上述 Actor–Critic-Identifier 架构，且目前大部分工作都集中在系统辨识器的设计上。需要注意的是，此类方法一般结构复杂，计算量较大，辨识器初始化阶段可能需要离线建模，“在线性”中等。不过，系统辨识器也可以作为一种生成模型生成一些额外的训练数据，因此其数据利用率和样本效率较高，这一点会在下一节进一步解释[104-105]。

4.2 离策略积分强化学习

这类方法无需对未知系统进行辨识，而是以数据驱动的方式直接学习值函数和控制律。2009 年，Vrabie 等人首先提出了积分强化学习（integral RL）的概念[106]，其核心思想是将 HJB 方程（14）转换成如下积分形式：

$$\begin{aligned} &V^{*}(x(t)) - V^{*}(x(t-T)) \\ &+\int_{t-T}^{t} U(x(\tau),u^{*}(\tau))d\tau = 0, \end{aligned} \tag{20}$$

其中，T 代表积分区间，$u^{*}(x) = -0.5R^{-1}g^{\top}\nabla V^{*}(x)$ 代表最优控制律。

相比于方程（14），积分形式的贝尔曼方程（20）同样该方法能够在不依赖系统未知动态 $F(x,u)$ 的前提下，通过在线数据采集来估计值函数并训练 Critic 网络。最初的积分强化学习仍需要离线训练，后续 Modares 和 Vamvoudakis 等人结合 SPI 算法提出了在线自适应版本的积分强化学习[107-108]。然而，由于求解最优控制律时仍然涉及未知的 $g(x)$，积分强化学习只能解决系统部分动态未知的最优控制问题[51]。

为了进一步实现完全不依赖系统动态信息的算法，Jiang 和 Luo 等人分别在 2014 年左右提出了离策略积分强化学习（off-policy integral RL）[109,110]，取得了重大突破。所谓的“离策略”是指智能体可以使用任意策略采集到的数据进行训练，从而将待学习的控制律 $\hat{u}$ 和实际作用于系统的控制律 u_s（也称作行为策略）相分离：

$$\dot{x} = f(x) + g(x)\hat{u} + g(x)(u_s - \hat{u}). \tag{21}$$

再利用 $2Ru^*(x) = -g^\top \nabla V^*(x)$ 这一重要性质和方程（20），即可推导如下形式的贝尔曼方程：

$$\begin{aligned} &\int_{t-T}^{t} \left(2(u^*)^\top R u_s - (u^*)^\top R u^* + S(x)\right) d\tau \\ &+V^*\left(x(t)\right) - V^*\left(x(t-T)\right) = 0 \end{aligned} \tag{22}$$

随后可以定义神经网络 $\hat{V}(x) = \hat{W}_c^\top \phi_c(x)$ 与 $\hat{u}(x) = \hat{W}_a^\top \phi_a(x)$，代入方程（22）后可以得到如下贝尔曼残差：

$$\begin{aligned} E(t) = &\hat{W}_c^\top \Delta\phi_c(t) + \mathrm{vec}\left(\hat{W}_a\right)^\top \eta(t) \\ &+ \mathrm{vec}\left(\hat{W}_a\right)^\top D(t)\,\mathrm{vec}\left(\hat{W}_a\right) + p(t) \end{aligned} \tag{23}$$

其中 $\Delta\phi_1(t) = \phi_1(t) - \phi_1(t-T)$，$\mathrm{vec}(\cdot)$ 代表向量化算子，$\eta(t)$、$D(t)$ 和 $p(t)$ 的具体形式可见文献[61,77]。

通过梯度法最小化式（23）即可训练 Actor-Critic 网络，

$$\begin{cases} \dot{\hat{W}}_c = -\alpha_1 \dfrac{\Delta\phi_c(t)}{m_s^2(t)} E(t) \\ \mathrm{vec}\left(\dot{\hat{W}}_a\right) = -\alpha_2 \dfrac{\eta(t)}{m_s^2(t)} E(t) \end{cases} \tag{24}$$

其中，α_1 和 α_2 为学习律，m_s 为正则化项。可以看到上述学习律完全不需要系统动态的先验知识，因此被后续很多工作广泛采用[77,111-113]。需要注意的是，离策略积分强化学习要求所采用的行为策略 u_s 能够稳定系统，这是使用该方法时必须满足的一个限制。在实际使用中，可能需要预先大规模仿真或历史数据离线训练获得安全稳定的行为策略。

4.3 Q 学习

Q 学习（或者 1.2.2 节提到的控制依赖启发式动态规划）作为一种经典的值函数法，在很早之前就被用来处理未知离散时间系统的控制问题[58,114-116]。然而，由于涉及求解偏微分形式的 HJB 方程（14），针对未知连续时间系统的 Q 学习算法直到最近几年才出现。尽管 Lee 和 Palanisamy 等人分别在 2012 年和 2015 年提出了基于积分强化学习的 Q 学习算法[117,118]，解决了未知连续时间线性系统的最优控制问题，但该方法只能离线迭代训练且需要初始稳定控制。2017 年，Vamvoudakis 等人成功提出了连续时间系统的在线 Q 学习控制方法[119]，且无需初始稳定控制。其基本思想是定义如下连续时间系统的 Q 函数：

$$\begin{aligned} Q(x,u) = &V^*(x) + U(x,u) \\ &+(\nabla V^*)^\top \left(f(x) + g(x)u\right). \end{aligned} \tag{25}$$

虽然式（25）中仍然涉及未知动态 $f(x)$ 和 $g(x)$，但通过 $Q(x,u^*) = V^*(x)$ 这一关键性质，可以借助积分贝尔曼方程(20）推导出以下 Q 函数形式的贝尔曼方程：

$$\begin{aligned} &Q(x(t),u^*(t)) - Q^*(x(t-T),u^*(t-T)) \\ &+\int_{t-T}^{t} U(x(\tau),u^*(\tau)) d\tau = 0, \end{aligned} \tag{26}$$

此时再使用 Actor-Critic 网络分别逼近 u^* 和 $Q(x,u)$，即可实现不依赖系统动态模型的在线更新。最近几年，一些研究进一步发展了这类 Q 学习算法，在连续时间非线性系统及时延系统的镇定控制、跟踪控制等方面取得了不错的效果[120-124]。值得一提的是，此类方法同样继承了 Q 学习的过估计、学习过程不稳定等缺点[46]。不过，其结构简单，不像前两个方法需要离线初始化准备，具有最强的“在线性”。

最后，在表 2 中对本节所梳理的三类方法进行对比总结。虽然这些方法在一定程度上突破了对系统精确动态模型的依赖，但仍存在各种各样的缺陷。特别地，在应用于工业控制的过程中，样本效率是影响强化学习实用性的瓶颈之一。为应对样本效率不足的问题，近年来提出了多种提升策略，主要包括：1）历史数据复用[125,126]：通过引入经验回放（replay buffer）或模仿学习等机制，提升已有数据的利用率，避免过度依赖新交互。2）系统模型辅助[100-101,104]：使用学习到的系统模型在模拟环境中生成样本，提高策略优化阶段的学习速度。3）迁移学习[127-129]：将预训练策略迁移到新任务中，或在任务间共享策略表示，提升泛化能力并减少数据需求。4）元学习[130-131]：使智能体在面对新任务时可通过少量样本迅速适应，已成为提升样本效率的重要研究方向。上述方法可降低对真实环境交互的依赖，为强化学习在实际工业控制系统中的部署提供有力支撑。

表 2　面向系统模型未知的在线强化学习方法总结

Tab.2　Summary of online RL methods for systems with unknown models

分类	算法核心	优势	限制	样本效率
基于模型的强化学习	Actor-Critic-Identifier 架构与系统辨识器设计	利用模型预测生成样本、样本效率高	结构复杂、计算量大	高
离策略积分强化学习	方程(22)与行为策略	可复用离线数据、收敛稳定	需要稳定的行为策略	中等
Q 学习	Q 函数(25)与方程（26）	实现简单、算法直观	过估计、学习过程不稳定	低

5　理论分析与性能边界

前几节系统地梳理了以强化学习为主的各种算法。然而，由于强化学习方法多数为数据驱动的“黑箱式优化”，其稳定性与收敛性有时会受到业界质疑。因此，本节主要针对上一节提到的一些经典在线强化学习方法，从 Lyapunov 稳定性、策略收敛性、前提假设等角度出发，逐步介绍现有方法的理论分析结果和性能边界，从而能够更好地理解并在实际使用它们。

在给出系统稳定性和算法收敛性分析之前，首先介绍一些必要的假设。

假设 1[61,132]：值函数 $V^*(x)$ 和最优控制律 $u^*(x)$ 在紧集 $x\in\Omega\subseteq\mathbb{R}^n$ 上可以表达为如下神经网络形式：

$$\begin{cases} V^*(x)=W_c^\top\phi_c(x)+\epsilon_c, \\ u^*(x)=W_a^\top\phi_a(x)+\epsilon_a, \end{cases} \tag{27}$$

其中，W_c 和 W_a 分别代表网络的理想权重，ϕ_c 和 ϕ_a 代表基函数，ϵ_c 和 ϵ_a 分别代表神经网络近似误差。同时，理想网络权重 W_c 和 W_a、基函数 ϕ_c 和 ϕ_a 以及近似误差 ϵ_c 和 ϵ_a 在紧集 Ω 上均有界。

上面的假设与著名的全局逼近定理有关[133]，在理论分析中可以保证网络参数能够收敛。值得注意的是该假设并不严苛，首先在一般情况下值函数和最优控制律都是连续可微的，满足适用全局逼近定理的前提条件。其次，基函数是由用户自己选择或设计的，因此总可以找到合适的函数保证其有界性（例如 Sigmoid 或 tanh 激活函数）。最后，假设 1 的成立需要保证系统的状态轨迹始终处于 Ω 内。Ω 的大小通常和基函数的激活范围有关，因此在一般情况下很难提前给出的估计，只能尽量保证系统的状态有界。当系统轨迹超出 Ω 后，公式（27）可能就不再成立。为缓解该问题，通常有两种做法：1）借助表征学习，通过结构化基函数设计（如 Fourier 基函数、径向基函数网络等），结合在线学习机制动态扩展 Ω 的覆盖范围[134-138]。2）采用安全强化学习，将系统状态约束在 Ω 内（详见第 6.2 节），以避免系统状态落入网络泛化能力较弱的区域[139-141]。

接下来的假设 2 对证明算法的收敛性也非常重要，该假设通常和信号的充分激励（persistently exciting）有关，但其具体形式因算法而异，这里以上一节提到的离策略积分强化学习为例。

假设 2[61,77]：式(23)中的信号 $\rho=[\Delta\phi_c^\top,\eta^\top]^\top$ 是充分激励的，即存在正常数 δ,β_1,β_2 使得下列关系在任意时刻 t 都成立：

$$\beta_1 I\le\int_t^{t+\delta}\rho(\tau)\rho^\top(\tau)\mathrm{d}\tau\le\beta_2 I. \tag{28}$$

该假设的主要作用是保证训练数据携带充分的信息，只有这样神经网络才能收敛。实际上，强化学习算法需要探索信号就是为了满足上述假设。不过，向系统注入探索信号来保证假设 2 成立的做法会或多或少地影响系统稳定性和控制性能，这算是强化学习在应用时的一大限制。为缓解该问题，通常有几种可行的办法：1）在学习律中引入经验回放[125]或同步学习（concurrent learning）[126]技术，从历史数据中筛选信息量高的样本，提高训练效率并减少实时激励需求。2）基于模型强化学习方法可以利用神经网络模型（19）生成含有足够激励的虚拟轨迹，用于辅助策略更新。3）课程学习[142-143]通过“由浅入深”的任务设计，引导智能体在安全可控的探索轨迹中逐步学习复杂行为，有效降低探索带来的安全风险。

最后，给出算法收敛性的分析结果。

定理 1[61,77]：令假设 1 和 2 成立，且行为策略 u_s 能够保证系统稳定，则离策略积分强化学习算法（24）可以保证神经网络估计误差 $\tilde{W}_c=W_c-\hat{W}_c$ 和 $\tilde{W}_a=W_a-\hat{W}_a$ 一致最终有界。

证明定理 1 需要先构造 Lyapunov 函数：$L=0.5\alpha_1^{-1}\tilde{W}_c^\top\tilde{W}_c+0.5\alpha_2^{-1}\mathrm{vec}(\tilde{W}_a)^\top\mathrm{vec}(\tilde{W}_a)$，随后沿着系统轨迹对 L 求导并利用 Lyapunov 扩展

定理完成证明，具体过程略。

对于上述结果需要进行两点说明。1）系统的稳定性依赖于提前设定好的行为策略。其实对于表 2 中提到的方法，总是需要对稳定性做出一定的前提假设。2）由于神经网络近似误差的存在，$\tilde{W}_c$ 和 $\tilde{W}_a$ 只能保证收敛到 0 的邻域，因此所得到的最终控制律只能是近似最优的，实际使用时可能会存在稳态误差。如何增强强化学习控制系统的稳定性和收敛性仍是亟待解决的难题，下一节会进一步讨论。

6 挑战与发展趋势

尽管近年来强化学习在最优控制问题中取得了显著进展，但其在实际工程系统中的部署仍面临诸多挑战。这些问题不仅源于强化学习本身在样本效率、泛化能力等方面的不足，更与控制系统对实时性、安全性和稳定性的严格要求密切相关。

相较于传统控制方法依赖精确系统模型的假设，强化学习以其数据驱动的优势，尤其适用于处理具有未知模型或黑箱结构的系统。然而，从前文综述可见，强化学习方法在实际应用中暴露出以下三大核心瓶颈：首先，状态信息不完备问题广泛存在于工业控制场景，而强化学习对状态数据的完整性要求极高。被誉为强化学习之父的 Richard Sutton 在其著作《强化学习导论》的最后指出，对完整状态信息的严格需求限制了大部分经典强化学习方法的应用[31]。其次，安全性保障不足是强化学习部署于物理系统时的关键限制因素。强化学习依赖试错，在训练过程中容易触发不可接受的危险行为。NASA 在其技术报告中明确指出："强化学习固有的问题是安全与探索之间的权衡" [144]，该问题在飞行控制、工业过程控制等任务中尤为突出。最后，稳定性与鲁棒性问题也成为强化学习控制可靠部署的重要阻碍。从上一节理论分析可以看出，强化学习的收敛性与稳定性严重依赖于初始稳定策略、数据丰富程度以及近似误差的影响，远未形成成熟可控的部署体系。

基于上述分析，本文将进一步围绕这三大挑战展开探讨，系统梳理当前强化学习在"状态信息不完备""安全保障机制不足""稳定性与鲁棒性差"等方面的主要瓶颈，并重点讨论以下代表性的发展趋势：1）信息不完备场景下的强化学习决策大模型；2）面向模型未知系统的安全强化学习；3）稳定性与鲁棒性增强的强化学习算法。

6.1 信息不完备场景下的强化学习决策大模型

前文所述的大多数强化学习方法普遍假设系统状态可完全观测，然而在实际工业控制场景中，这一假设常常难以满足。例如，在桥式起重机等典型工业装备中，系统全状态通常包括台车的位置与速度、负载摆角及其角速度。然而由于成本和传感器部署限制，诸如摆角角速度等状态变量往往不可直接测量[145-147]。在强化学习领域，对于这类信息不完备系统的控制问题，通常采用部分可测马尔可夫决策过程（Partially Observable Markov Decision Process，POMDP）进行建模[148-149]。POMDP 和一般的 MDP 最大区别在于引入了观测空间 O。正是由于 POMDP 中的观测不再满足马尔可夫性，策略学习的复杂度显著提升。

在此情形下，有效的控制器应具备一定的时序记忆能力，能够通过历史观测序列隐式或显式地恢复系统的潜在状态信息。传统的方法按照控制理论的思路，通过引入观测器对系统状态进行估计[148,150]。近年来，有学者尝试将具备时序建模能力的大模型结构引入强化学习框架，尤其是通过引入 Transformer 或变分自编码器等模块，从历史观测序列中高效提取状态信息，从而解决部分状态可测问题[16-17]。

在机器人控制、灵巧操作、长时序规划等任务中，这一类基于大模型的强化学习方法已取得初步进展[151]。例如，Decision Transformer 借鉴自然语言处理中的条件生成建模思想，通过对观测、动作和奖励组成的轨迹序列进行建模，实现了不显式依赖值函数或 Actor 网络的动作预测，代表了从"结构显式策略"向"序列建模策略"的范式转变[152]。另一方面，Trajectory Transformer 利用 Transformer 结构建模历史观测轨迹分布，从中抽样未来动作序列，兼具良好的样本效率与泛化能力[153]。这些方法拓展了传统强化学习在控制问题中的范式，使得策略优化提升至轨迹级的全局优化，为应对复杂环境下的智能控制提供了新工具。

尽管如此，当前基于大模型的强化学习方法在工业部署中仍面临诸多挑战：1）模型参数规模庞大，推理计算开销高；2）策略生成过程缺乏可解释性，难以满足工程应用对安全性与稳定性的高要求；3）现有方法多依赖离线训练，尚未具备在线实时适应能力。因此，未来的研究需将控制理论中的稳定性、安全性分析工具嵌入决策大模型强化学习框架，采用轻量化大模型设计或在线微调与增量学习结合等方式，从结构设计、训练机制、性能保障等多个方面提升其在在线控制中的模型效率与部署可行性，推动强化学习方法在信息不完备的高维复杂系统中实用化落地。

6.2 面向模型未知系统的安全强化学习

前述内容主要关注了强化学习算法对系统稳定性与性能最优化的能力。然而，在工业 4.0 时

代，稳定性只是现代工业系统的基本要求，如何在不确定环境下实现安全与最优的双重目标，已成为智能制造与过程控制中的核心挑战。所谓安全性，通常是指系统状态 x 需始终保持在某个安全工作域内（$h(x)\geq 0$），例如在某些化工过程中，尽管温度需调节至设定值，但其波动范围也必须受到严格限制，以避免安全事故。通常，这类问题在数学上可以描述为受限最优控制（constrained optimal control）问题。

对于系统动态模型未知的情形，直接求解受限最优控制问题困难重重。一方面，实际系统本身可能具有非线性、滞后与惯性，再加上动态模型未知，导致控制器无法对未来状态进行准确预测，也就无法提前规避风险。另一方面，强化学习采用基于“试错”的策略优化机制，其探索过程本身可能触发不安全行为，尤其在早期学习阶段或策略切换过程中，这类安全风险尤为突出。

为应对上述挑战，近年来出现了多种面向模型未知系统的安全强化学习方法。其目标是在策略优化过程中对系统状态进行有效约束，确保控制策略的安全可部署性。一种常见思路是构建软约束优化问题，即在效用函数中添加与状态约束相关的惩罚项，使得强化学习倾向于生成满足约束的策略。这类方法通常基于概率强化学习框架，其效用函数可表述为如下形式[139-141]：

$$U(x,u)=x^{\top}Sx+u^{\top}Ru-\log\frac{\gamma h(x)}{\gamma h(x)+1} \quad (29)$$

其中前两项为普通的二次型性能指标，而最后一项与系统约束 $h(x)\geq 0$ 有关。当系统状态将要违反约束时，U 会趋向于无穷大，从而迫使系统回到安全运行范围内。这类方法实现简单，但本质上仅能“降低”违约风险，无法提供严格的安全性保证。

为实现更强的安全性保障，基于控制障碍函数的强化学习方法近年来受到广泛关注。控制障碍函数利用前向不变性理论，将状态约束转化为控制输入约束，从而构造一个安全可行域，确保系统轨迹始终处于该域内。在数据驱动场景下，这类方法通常与其他算法联合使用，由强化学习 Actor 网络的输出作为候选动作，再由控制障碍函数设计的安全滤波器进行修正。这种架构有效结合了强化学习与控制理论的安全保障能力，已在自动驾驶、机器人等任务中取得良好效果[154-156]。最后，基于松弛函数的方法可以将受限最优控制问题转换为增广系统的无约束最优控制问题，从而避免直接处理状态约束[157-158]。当系统模型未知时，可以对增广系统进行数据驱动建模，再结合基于模型强化学习构建安全控制器[159]。

尽管这些方法各具优势，当前安全强化学习仍面临诸多挑战，例如如何设计收敛性强、稳定性好的策略更新机制，以及如何应对现实系统中高维状态空间下的安全域表达与控制可行性问题。针对这些问题，未来可考虑引入不确定性感知建模方法，从而进一步推动安全强化学习在复杂工业系统的应用。

6.3 稳定性与鲁棒性增强的强化学习算法

虽然安全强化学习方法能够在受限约束下提升系统的运行安全性，但其前提往往是系统已有稳定的控制器或足够的策略初始化保障。而在实际应用中，强化学习从零开始学习策略时，其对系统稳定性的需求常常无法满足，特别是在模型未知或扰动频繁的工业控制场景下。因此，有必要进一步探讨如何增强强化学习的稳定性和鲁棒性，以提升其工程部署能力。这一问题正是本节将重点讨论的内容。

目前，大多数强化学习算法都隐含地依赖“初始稳定/容许控制策略”的假设，即要求 Actor 网络在训练初期就能生成一个使系统稳定的控制输入。然而，在系统动态模型未知的情形下，这一假设往往难以满足。更进一步地，即使训练过程取得了较好的性能提升，控制策略往往仍容易在环境发生微小扰动时出现性能退化甚至不稳定现象，表现出较差的泛化能力。因此，如何在模型未知条件下提升强化学习策略的稳定性和鲁棒性，是当前智能控制研究的核心问题之一。

为了解决初始化依赖的问题，部分研究工作尝试从 Lyapunov 函数入手，构建无需初始稳定控制器的强化学习算法。对于模型未知的线性系统，有研究提出了基于值迭代的方案[160-161]其核心思想是先构造一个候选的正定 Lyapunov 函数：$V=x^{\top}P^{0}x$。随后，按照如下方式迭代更新即可找到最优值函数 P^{*}：

$$\begin{cases}\int_{t-T}^{t}x^{\top}(\tau)H^{j}x(\tau)-2(Ru(\tau))^{\top}K^{j}x(\tau)\mathrm{d}\tau \\ =x^{\top}(t)P^{j}x(t)-x^{\top}(t-T)P^{j}x(t-T) \\ P^{j+1}=P^{j}+\epsilon_{j}(H^{j}+S-(K^{j})^{\top}RK^{j})\end{cases} \quad (30)$$

相比于寻找初始稳定的增益 K^{0}，确定一个初始的正定矩阵 P^{0} 更加容易，因而可有效放宽算法的应用前提。不过，该类方法收敛速度较慢，对系统噪声和建模误差也较为敏感。

在非线性系统中，研究者进一步发展出 Lyapunov 引导的强化学习网络更新机制[29,159]。这类方法通常引入一个辅助 Lyapunov 函数，并沿负梯度方向更新网络权重，从而在学习过程中持续维持系统闭环稳定性。这种设计不仅能显著降低训练风险，也为后续安全性保障打下基础。此外，也有方法尝试将强化学习与其他控制理论

方法结合，以实现更可靠的控制性能。例如在最优跟踪控制任务中，可以仅使用强化学习作为轨迹优化器，而将低层跟踪任务交由自适应控制器完成[26]。这种二自由度的架构充分利用了强化学习的暂态优化能力与自适应控制的渐进稳定性，使系统能够兼顾暂态性能与稳态精度，类似框架可参考文献[162-163]。

针对鲁棒性的提升，除了传统基于 H 无穷等鲁棒控制设计的框架[71]，近年来结合人工智能领域的域随机化（domain randomization）技术成为研究热点之一[164-165]。该方法的思路是通过在训练过程中反复扰动系统的物理参数、噪声水平或任务目标，使智能体在多样化的环境中习得泛化能力更强的策略。然而，使用域随机化往往要求具备高保真度的仿真器支持，并伴随较长的训练周期。除此之外，迁移强化学习（transfer RL）也被广泛用于提升控制器在新场景、新系统中的适应能力[127-129]。需要注意的是，和传统迁移学习不同，控制器的迁移需要算法能够学习和控制律耦合的动态映射[129,166]，因此通常需要设计复杂的对齐机制，同时对训练数据的获取提出了更高的要求。

总体来看，当前在线强化学习方法在实际中面临状态不完备、安全性不足和稳定性鲁棒性差等挑战，显著制约了强化学习在工业场景中的部署效果。本节总结了为应对上述问题而发展出的关键技术路径，为今后在高性能、高安全、高稳定性的强化学习发展方向提供参考。

7 结论

在工业过程控制、航空航天、机器人等高端装备系统中，系统动态模型未知、参数时变、关键变量无法直接测量以及外部扰动频发是普遍存在的现实问题，使得传统依赖精确数学建模的控制方法面临严峻挑战。强化学习作为一种典型的数据驱动控制方法，能够通过与环境交互持续优化控制策略，在应对模型未知条件下的最优控制任务中展现出显著优势。本文围绕连续时间系统中的动态模型未知问题，系统回顾了通用强化学习的基础理论、面向模型已知场景最优控制的强化学习方法、适用于模型未知场景的在线强化学习方法、相关的理论分析工具以及工业落地中的实际挑战。通过对工程实例中模型未知问题的梳理以及现有方法的理论分析结果可以看出，尽管该领域研究不断推进，当前方法仍在状态信息不完备、安全性以及闭环稳定性保障等方面存在不足。未来仍需进一步融合控制理论与人工智能技术，提升强化学习策略的可解释性、安全性和稳定性，构建面向流程工业、智能制造等典型实际系统的强化学习智能控制框架，从而推动强化学习在复杂动态系统中的落地与广泛应用。

参考文献

[1] 王喜文. 工业 4.0、互联网+、中国制造 2025 中国制造业转型升级的未来方向[J]. 国家治理，2015(23): 12-19.
WANG X W. Industry 4.0, internet plus, made in China 2025: future direction of transformation and upgrading of China's manufacturing industry[J]. National Governance, 2015(23): 12-19. (in Chinese)

[2] Research Group of Institute of Industrial Economics. New industrialization and the development direction of China's manufacturing industry during the 14th five-year plan period[J]. China Economist, 2020, 15(4): 38-63. (in Chinese)

[3] 赵春晖，余万科，柴铮，等. 运行工况监测与故障溯源推理：机器学习方法[M]. 北京：化学工业出版社，2022.
ZHAO C H, YU W K, CHAI Z, et al. Operating condition monitoring and fault tracing reasoning[M]. Beijing: Chemical Industry Press, 2022. (in Chinese)

[4] 万百五，韩崇昭，蔡远利. 控制论——概念、方法与应用[M]. 北京：清华大学出版社, 2009.
WAN B W, HAN C Z, CAI Y L. Cybernetics: concepts methods and applications[M]. Beijing: Tsinghua University Press, 2009. (in Chinese)

[5] ZHAO C H. Perspectives on nonstationary process monitoring in the era of industrial artificial intelligence[J]. Journal of Process Control, 2022, 116: 255-272.

[6] 托马斯·瑞德. 机器崛起：遗失的控制论历史[M]. 王晓，郑心湖，王飞跃，译. 北京：机械工业出版社，2017.
RID T. Rise of the machines: a cybernetics history [M]. WANG X, ZHENG X H, WANG F Y, translate. Beijing: China Machine Press, 2017. (in Chinese)

[7] 喻林. 基于强化学习的连续非线性系统数据驱动控制问题研究 [D]. 合肥：中国科学技术大学，2024.
YU L. Research on data-driven control of continuous nonlinear systems based on reinforcement learning [D]. Hefei: University of Science and Technology of China, 2024. (in Chinese)

[8] 解学书. 最优控制理论与应用[M]. 北京: 清华大学出版社, 1986.
XIE X S. Optimal control theory and application[M]. Beijing: Tsinghua University Press, 1986. (in Chinese).

[9] DYDEK Z T, ANNASWAMY A M, LAVRETSKY E. Adaptive control and the NASA X-15-3 flight revisited[J]. IEEE Control Systems Magazine, 2010, 30(3): 32-48.

[10] STEIN G. Respect the unstable [J]. IEEE Control Systems Magazine, 2003, 23(4): 12-25.

[11] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of Go with deep neural networks and tree search[J]. Nature, 2016, 529(7587): 484-489.

[12] 刘源, 王初. 开物成务: 2024 年诺贝尔化学奖“计算蛋白质设计与结构预测”[J]. 中国科学基金, 2024, 38(6): 994-996.
LIU Y, WANG C. Unlocking nature's secrets: the 2024 Nobel Prize in chemistry for“computational protein design and structure prediction” [J]. Chinese Science Foundation, 2024, 38(6): 994-996. (in Chinese)

[13] DEEPSEEK-AI, GUO D Y, YANG D J, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning [EB/OL]. (2025-01-22) [2025-03-20]. https://arxiv.org/abs/2501.12948.

[14] 王鼎, 赵明明, 刘德荣, 等. 数据驱动自适应评判控制研究进展[J]. 自动化学报, 2025, 51(6): 1170-1190.
WANG D, ZHAO M M, LIU D R, et al. Research progress on data-driven adaptive critic control[J]. Acta Automatica Sinica, 2025, 51(6): 1170-1190. (in Chinese)

[15] ZHANG Y H, ZOU L, LIU Y, et al. A brief survey on nonlinear control using adaptive dynamic programming under engineering-oriented complexities[J]. International Journal of Systems Science, 2023, 54(8): 1855-1872.

[16] FIGUEIREDO PRUDENCIO R, MAXIMO M R O A, COLOMBINI E L. A survey on offline reinforcement learning: taxonomy, review, and open problems[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(8): 10237-10257.

[17] CAO Y J, ZHAO H, CHENG Y H, et al. Survey on large language model-enhanced reinforcement learning: concept, taxonomy, and methods[J]. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(6): 9737-9757.

[18] REN C, DING Y T, HU L, et al. Active disturbance rejection control of Euler-Lagrange systems exploiting internal damping[J]. IEEE Transactions on Cybernetics, 2022, 52(6): 4334-4345.

[19] CHAI Z, ZHAO C H, HUANG B. Cross-domain knowledge transfer in industrial process monitoring: a survey[J]. Journal of Process Control, 2025, 149: 103408.

[20] SMITH C A, CORRIPIO A B. Principles and practice of automatic process control[M]. 3rd ed. New York: John Wiley & Sons, 2005.

[21] ZHANG H R, ZHAO C H. Residual integral inverse reinforcement learning for intelligent self-healing control of unknown systems with actuator faults[J]. Nonlinear Dynamics, 2025, 113(2): 1353-1369.

[22] LEWIS F L, VRABIE D L, SYRMOS V L. Optimal control[M]. 3rd ed. Hoboken, New Jersy: John Wiley & Sons, 2012.

[23] ÅSTRÖM K J, MURRAY R M. Feedback systems: an introduction for scientists and engineers[M]. Princeton, New Jersy: Princeton University Press, 2021.

[24] BHATTACHARYYA S P. Robust control under parametric uncertainty: an overview and recent results[J]. Annual Reviews in Control, 2017, 44: 45-77.

[25] NGUYEN N T. Model-reference adaptive control: a primer[M]. Cham, Switzerland: Springer International Publishing, 2018.

[26] ZHANG H R, ZHAO C H, DING J L. Constrained reinforcement learning-based closed-loop reference model for optimal tracking control of unknown continuous-time systems[J]. IEEE Transactions on Automation Science and Engineering, 2024, 21(4): 7312-7324.

[27] FARRELL J A, POLYCARPOU M M. Adaptive approximation based control: unifying neural, fuzzy and traditional adaptive approximation approaches[M]. New York: John Wiley & Sons, Inc., 2006.

[28] CHOPRA N, FUJITA M, ORTEGA R, et al. Passivity-based control of robots: theory and examples from the literature[J]. IEEE Control Systems Magazine, 2022, 42(2): 63-73.

[29] ZHANG H R, ZHAO C H, DING J L. Online reinforcement learning with passivity-based stabilizing term for real time overhead crane control without knowledge of the system model[J]. Control Engineering Practice, 2022, 127: 105302.

[30] 侯忠生. 无模型自适应控制的现状与展望[J]. 控制理论与应用, 2014, 23(4): 586-592.
HOU Z S. On model-free adaptive control: the state of the art and perspective[J]. Control Theory & Applications, 2014, 23(4): 586-592. (in Chinese)

[31] SUTTON R S, BARTO A G. Reinforcement learning: an introduction [M]. 2nd ed. Bradford: Bradford Books, 2018.

[32] 刘全, 翟建伟, 章宗长, 等. 深度强化学习综述[J]. 计算机学报, 2018, 41(1): 1-27.
LIU Q, ZHAI J W, ZHANG Z C, et al. A survey of deep reinforcement learning[J]. Chinese Journal of Computers, 2018, 41(1): 1-27. (in Chinese)

[33] GRONDMAN I, BUSONIU L, LOPES G A D, et al. A survey of actor-critic reinforcement learning: standard and natural policy gradients[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2012, 42(6): 1291-1307.

[34] WATKINS C. Learning from delayed rewards[D]. London: King's College, 1989.

[35] HASSELT H V. Double Q-learning[C]//Proceedings of the Advances in Neural Information Processing Systems, 2010: 2613-2621.

[36] WILLIAMS R J. Simple statistical gradient-following algorithms for connectionist reinforcement learning[J]. Machine Learning, 1992, 8: 229-256.

[37] SUTTON R S, MCALLESTER D A, SINGH S P, et al. Policy gradient methods for reinforcement learning with function approximation[C]//Proceedings of the 13th International Conference on Neural Information Processing Systems, 1999: 1057–1063.

[38] BHATNAGAR S, SUTTON R S, GHAVAMZADEH M, et al. Natural actor–critic algorithms[J]. Automatica, 2009, 45(11): 2471-2482.

[39] 李敏. 针对连续动作控制的深度强化学习算法研究[D]. 成都: 电子科技大学, 2023.
LI M. Research on deep reinforcement learning algorithm for continuous action control[D]. Chengdu: University of Electronic Science and Technology of China, 2023. (in Chinese)

[40] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

[41] VAN HASSELT H, GUEZ A, SILVER D. Deep reinforcement learning with double Q-learning[C]//Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix, USA, 2016: 2094-2100.

[42] WANG Z, SCHAUL T, HESSEL M, et al. Dueling network architectures for deep reinforcement learning[C]//Proceedings of the 33rd International Conference on Machine Learning, 2016: 1995-2003.

[43] HESSEL M, MODAYIL J, VAN HASSELT H, et al. Rainbow: combining improvements in deep reinforcement learning[C]//Proceedings of the 32th AAAI Conference on Artificial Intelligence, 2018: 3215-3222.

[44] SCHULMAN J, LEVINE S, MORITZ P, et al. Trust region policy optimization[C]//Proceedings of the 32nd International Conference on Machine Learning, 2015: 1889-1897.

[45] SILVER D, LEVER G, HEESS N, et al. Deterministic policy gradient algorithms[C]//Proceedings of the 31st International Conference on Machine Learning. Beijing, China, 2014: 387-395.

[46] FUJIMOTO S, VAN HOOF H, MEGER D. Addressing function approximation error in actor-critic methods[C]//Proceedings of the 35th International Conference on Machine Learning, 2018: 1587-1596.

[47] HAARNOJA T, ZHOU A, ABBEEL P, et al. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor[C]//Proceedings of the 35th International Conference on Machine Learning, 2018: 1861-1870.

[48] 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[J]. 自动化学报, 2020, 46(7): 1301-1312.
SUN C Y, MU C X. Important scientific problems of multi-agent deep reinforcement learning[J]. Acta Automatica Sinica, 2020, 46(7): 1301-1312. (in Chinese)

[49] OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback[C]//Proceedings of the 36th Conference on Neural Information Processing Systems, 2022: 27730-27744.

[50] 张化光, 张欣, 罗艳红, 等. 自适应动态规划综述. 自动化学报, 2013, 39(4): 303-311.
ZHANG H G, ZHANG X, LUO Y H, et al. An overview of research on adaptive dynamic programming: an overview of research on adaptive dynamic programming[J]. Acta Automatica Sinica, 2013, 39(4): 303-311.

[51] LEWIS F L, VRABIE D, VAMVOUDAKIS K G. Reinforcement learning and feedback control: using natural decision methods to design optimal adaptive controllers[J]. IEEE Control Systems Magazine, 2012, 32(6): 76-105.

[52] RECHT B. A tour of reinforcement learning: the view from continuous control[J]. Annual Review of Control, Robotics, and Autonomous Systems, 2019, 2: 253-279.

[53] BUŞONIU L, DE BRUIN T, TOLIĆ D, et al. Reinforcement learning for control: performance, stability, and deep approximators[J]. Annual Reviews in Control, 2018, 46: 8-28.

[54] XU X, ZUO L, HUANG Z H. Reinforcement learning algorithms with function approximation: Recent advances and applications[J]. Information Sciences, 2014, 261: 1-31.

[55] MURRAY J J, COX C J, LENDARIS G G, et al. Adaptive dynamic programming[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2002, 32(2): 140-153.

[56] WERBOS P J. Consistency of HDP applied to a simple reinforcement learning problem[J]. Neural Networks, 1990, 3(2): 179-189.

[57] WERBOS P J. Approximate dynamic programming for real-time control and neural modelling[M]//Handbook of intelligent control. New York, USA: Van Nostrand Reinhold, 1992: 493-525.

[58] SI J, WANG Y T. Online learning control by association and reinforcement[J]. IEEE Transactions on Neural Networks, 2001, 12(2): 264-276.

[59] PROKHOROV D V, WUNSCH D C. Adaptive critic designs[J]. IEEE Transactions on Neural Networks, 1997, 8(5): 997-1007.

[60] PADHI R, UNNIKRISHNAN N, WANG X H, et al. A single network adaptive critic (SNAC) architecture for optimal control synthesis for a class of nonlinear systems[J]. Neural Networks, 2006, 19(10): 1648-1660.

[61] ZHU Y H, ZHAO D B. Comprehensive comparison of online ADP algorithms for continuous-time optimal control[J]. Artificial Intelligence Review, 2018, 49(4): 531-547.

[62] BEARD R W, SARIDIS G N, WEN J T. Galerkin approximations of the generalized Hamilton-jacobi-bellman equation[J]. Automatica, 1997, 33(12): 2159-2177.

[63] ABU-KHALAF M, LEWIS F L. Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network HJB approach[J]. Automatica, 2005, 41(5): 779-791.

[64] VAMVOUDAKIS K G, LEWIS F L. Online actor–critic algorithm to solve the continuous-time infinite horizon optimal control problem[J]. Automatica, 2010, 46(5): 878-888.

[65] WALLACE B A, SI J. Continuous-time reinforcement learning control: a review of theoretical results, insights on performance, and needs for new designs[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(8): 10199-10219.

[66] LIU D R, XUE S, ZHAO B, et al. Adaptive dynamic programming for control: a survey and recent advances[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2021, 51(1): 142-160.

[67] LIU D R, YANG X, WANG D, et al. Reinforcement-learning-based robust controller design for continuous-time uncertain nonlinear systems subject to input constraints[J]. IEEE Transactions on Cybernetics, 2015, 45(7): 1372-1385.

[68] WANG D, LIU D R, ZHANG Q C, et al. Data-based adaptive critic designs for nonlinear robust optimal control with uncertain dynamics[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2016, 46(11): 1544-1555.

[69] WANG D, MU C X. A novel neural optimal control framework with nonlinear dynamics: Closed-loop stability and simulation verification[J]. Neurocomputing, 2017, 266: 353-360.

[70] XUE S, LUO B, LIU D R, et al. Adaptive dynamic programming based event-triggered control for unknown continuous-time nonlinear systems with input constraints[J]. Neurocomputing, 2020, 396: 191-200.

[71] LI J, NAGAMUNE R, ZHANG Y H, et al. Robust approximate dynamic programming for nonlinear systems with both model error and external disturbance[J]. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(1): 896-910.

[72] KARIMI-GHARTEMANI M, ALI KHAJEHODDIN S, JAIN P, et al. Linear quadratic output tracking and disturbance rejection[J]. International Journal of Control, 2011, 84(8): 1442-1449.

[73] ANDERSON B D O, MOORE J B. Optimal control: linear quadratic methods[M]. Englewood Cliffs, USA: Prentice-Hall, Inc., 1989.

[74] ZHANG H G, CUI L L, ZHANG X, et al. Data-driven robust approximate optimal tracking control for unknown general nonlinear systems using adaptive dynamic programming method[J]. IEEE Transactions on Neural Networks, 2011, 22(12): 2226-2236.

[75] MODARES H, LEWIS F L. Linear quadratic tracking control of partially-unknown continuous-time systems using reinforcement learning[J]. IEEE Transactions on Automatic Control, 2014, 59(11): 3051-3056.

[76] MODARES H, LEWIS F L. Optimal tracking control of nonlinear partially-unknown constrained-input systems using integral reinforcement learning[J]. Automatica, 2014, 50(7): 1780-1792.

[77] ZHU Y H, ZHAO D B, LI X J. Using reinforcement learning techniques to solve continuous-time non-linear optimal tracking problem without system dynamics[J]. IET Control Theory & Applications, 2016, 10(12): 1339-1347.

[78] ZHAO J G. Neural network-based optimal tracking control of continuous-time uncertain nonlinear system via reinforcement learning[J]. Neural Processing Letters, 2020, 51(3): 2513-2530.

[79] ROSHANRAVAN S, SHAMAGHDARI S. Adaptive fault-tolerant tracking control for affine nonlinear systems with unknown dynamics via reinforcement learning[J]. IEEE Transactions on Automation Science and Engineering, 2024, 21(1): 569-580.

[80] ZHAO J, NA J, GAO G B. Robust tracking control of uncertain nonlinear systems with adaptive dynamic programming[J]. Neurocomputing, 2022, 471: 21-30.

[81] YANG X, LIU D R, WEI Q L, et al. Guaranteed cost neural tracking control for a class of uncertain nonlinear systems using adaptive dynamic programming[J]. Neurocomputing, 2016, 198: 80-90.

[82] 赵建国. 多时间尺度互联系统强化学习最优跟踪控制与应用[D]. 徐州: 中国矿业大学, 2023.
ZHAO J G. Reinforcement learning and optimal tracking control of multi-time-scale interconnected systems with applications[D]. Xuzhou: China University of Mining and Technology, 2023. (in Chinese)

[83] NA J, LYU Y F, ZHANG K Q, et al. Adaptive identifier-critic-based optimal tracking control for nonlinear systems with experimental validation[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2022, 52(1): 459-472.

[84] LI C, DING J L, LEWIS F L, et al. A novel adaptive dynamic programming based on tracking error for nonlinear discrete-time systems[J]. Automatica, 2021, 129: 109687.

[85] WANG D, ZHAO H L, ZHAO M M, et al. Novel optimal trajectory tracking for nonlinear affine systems with an advanced critic learning structure[J]. Neural Networks, 2022, 154: 131-140.

[86] SERESHKI Z T, ALI TALEBI H, ABDOLLAHI F. An analytical adaptive optimal control approach without solving HJB equation for nonlinear systems with input constraints[J]. IET Control Theory & Applications, 2024, 18(10): 1275-1288.

[87] SERESHKI Z T, TALEBI H A, ABDOLLAHI F. A nonlinear adaptive H∞ optimal control method without solving HJIE: an analytical approach[J]. IEEE Transactions on Automatic Control, 2024, 69(6): 4126-4133.

[88] ZHAO J G, YANG C Y, GAO W N, et al. Optimal dynamic controller design for linear quadratic tracking problems[J]. IEEE Transactions on Automatic Control, 2024, 69(6): 4021-4027.

[89] ZHAO J G, YANG C Y, GAO W N, et al. Incremental reinforcement learning and optimal output regulation under unmeasurable disturbances[J]. Automatica, 2024, 160: 111468.

[90] XIE K D, YU X, LAN W Y. Optimal output regulation for unknown continuous-time linear systems by internal model and adaptive dynamic programming[J]. Automatica, 2022, 146: 110564.

[91] AMIRPARAST A, HOSSEINI SANI S K. Undiscounted reinforcement learning for infinite-time optimal output tracking and disturbance rejection of discrete-time LTI systems with unknown dynamics[J]. International Journal of Systems Science, 2023, 54(10): 2175-2195.

[92] CHEN C, MODARES H, XIE K, et al. Reinforcement learning-based adaptive optimal exponential tracking control of linear systems with unknown dynamics[J]. IEEE Transactions on Automatic Control, 2019, 64(11): 4423-4438.

[93] GAO W N, JIANG Z P. Learning-based adaptive optimal output regulation of linear and nonlinear systems: an overview[J]. Control Theory and Technology, 2022, 20(1): 1-19.

[94] CHEN J W, ZHAO C H. Addressing information asymmetry: deep temporal causality discovery for mixed time series[J]. IEEE Transactions on Pattern

Analysis and Machine Intelligence, 2025, 47(7): 5723-5741.

[95] 李想. 基于强化学习的柔性协作机器人控制研究[D]. 合肥: 中国科学技术大学, 2024.
LI X. Research on flexible collaborative robot control based on reinforcement learning[D]. Hefei: University of Science and Technology of China, 2024. (in Chinese)

[96] RAVICHANDAR H, POLYDOROS A S, CHERNOVA S, et al. Recent advances in robot learning from demonstration[J]. Annual Review of Control, Robotics, and Autonomous Systems, 2020, 3: 297-330.

[97] 石静. 基于自适应动态规划的非线性系统最优控制及其在微电网中的应用[D]. 南京: 南京邮电大学, 2020.
SHI J. Research on flexible collaborative robot control based on reinforcement learning[D]. Nanjing: Nanjing University of Posts and Telecommunications, 2020. (in Chinese)

[98] FU Y, LI B, FU J. Multi-model adaptive switching control of a nonlinear system and its applications in a smelting process of fused magnesia[J]. Journal of Process Control, 2022, 115: 67-76.

[99] 柴铮. 面向工业监控典型欠数据场景的知识迁移方法研究[D]. 杭州: 浙江大学, 2022.
CHAI Z. Knowledge transfer methods for typical industrial monitoring scenarios with limited data[D]. Hangzhou: Zhejiang University, 2022. (in Chinese)

[100] BHASIN S, KAMALAPURKAR R, JOHNSON M, et al. A novel actor–critic–identifier architecture for approximate optimal control of uncertain nonlinear systems[J]. Automatica, 2013, 49(1): 82-92.

[101] MODARES H, LEWIS F L, NAGHIBI-SISTANI M B. Adaptive optimal control of unknown constrained-input systems using policy iteration and neural networks[J]. IEEE Transactions on Neural Networks and Learning Systems, 2013, 24(10): 1513-1525.

[102] LYU Y F, NA J, YANG Q M, et al. Online adaptive optimal control for continuous-time nonlinear systems with completely unknown dynamics[J]. International Journal of Control, 2016, 89(1): 99-112.

[103] ZHAO B, LIU D R, LUO C M. Reinforcement learning-based optimal stabilization for unknown nonlinear systems subject to inputs with uncertain constraints[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(10): 4330-4340.

[104] KAMALAPURKAR R, WALTERS P, DIXON W E. Model-based reinforcement learning for approximate optimal regulation[J]. Automatica, 2016, 64: 94-104.

[105] JIA C X, ZHANG F X, XU T, et al. Model gradient: unified model and policy learning in model-based reinforcement learning[J]. Frontiers of Computer Science, 2023, 18(4): 184339.

[106] VRABIE D, LEWIS F. Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems[J]. Neural Networks, 2009, 22(3): 237-246.

[107] MODARES H, LEWIS F L, NAGHIBI-SISTANI M B. Integral reinforcement learning and experience replay for adaptive optimal control of partially-unknown constrained-input continuous-time systems[J]. Automatica, 2014, 50(1): 193-202.

[108] VAMVOUDAKIS K G, VRABIE D, LEWIS F L. Online adaptive algorithm for optimal control with integral reinforcement learning: online adaptive algorithm for optimal control[J]. International Journal of Robust and Nonlinear Control, 2014, 24(17): 2686-2710.

[109] JIANG Y, JIANG Z P. Robust adaptive dynamic programming and feedback stabilization of nonlinear systems[J]. IEEE Transactions on Neural Networks and Learning Systems, 2014, 25(5): 882-893.

[110] LUO B, WU H N, HUANG T W, et al. Data-based approximate policy iteration for affine nonlinear continuous-time optimal control design[J]. Automatica, 2014, 50(12): 3281-3290.

[111] SONG R Z, LEWIS F L, WEI Q L, et al. Off-policy actor-critic structure for optimal control of unknown systems with disturbances[J]. IEEE Transactions on Cybernetics, 2016, 46(5): 1041-1050.

[112] LI J N, CHAI T Y, LEWIS F L, et al. Off-policy interleaved Q-learning: optimal control for affine nonlinear discrete-time systems[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(5): 1308-1320.

[113] GUO L, ZHAO H. Online adaptive optimal control algorithm based on synchronous integral reinforcement learning with explorations[J]. Neurocomputing, 2023, 520: 250-261.

[114] LIU F, SUN J, SI J, et al. A boundedness result for the direct heuristic dynamic programming[J]. Neural Networks, 2012, 32: 229-235.

[115] SOKOLOV Y, KOZMA R, WERBOS L D, et al. Complete stability analysis of a heuristic approximate dynamic programming control design[J]. Automatica, 2015, 59: 9-18.

[116] ZHAO D B, XIA Z P, WANG D. Model-free optimal control for affine nonlinear systems with convergence analysis[J]. IEEE Transactions on Automation Science and Engineering, 2015, 12(4): 1461-1468.

[117] LEE J Y, PARK J B, CHOI Y H. Integral Q-learning and explorized policy iteration for adaptive optimal control of continuous-time linear systems[J]. Automatica, 2012, 48(11): 2850-2859.

[118] PALANISAMY M, MODARES H, LEWIS F L, et al. Continuous-time Q-learning for infinite-horizon discounted cost linear quadratic regulator problems[J]. IEEE Transactions on Cybernetics, 2015, 45(2): 165-176.

[119] VAMVOUDAKIS K G. Q-learning for continuous-time linear systems: a model-free infinite horizon optimal control approach[J]. Systems & Control Letters, 2017, 100: 14-20.

[120] CHEN A S, HERRMANN G. Adaptive optimal control via continuous-time Q-learning for unknown nonlinear affine systems[C]//Proceedings of the IEEE 58th Conference on Decision and Control (CDC), 2019: 1007-1012.

[121] BENHMIDOUCH Z, ALAOUI S B, ABBOU A, et al. A Q-learning approach to model-free infinite horizon control for linear time delay systems[C]//Proceedings of the IEEE 63rd Conference on Decision and Control (CDC), 2024: 3335-3340.

[122] CHEN A S, HERRMANN G. An adaptive critic learning approach for nonlinear optimal control subject to excitation and weight constraints[C]//Proceedings of the 62nd IEEE Conference on Decision and Control (CDC), 2023: 2497-2502.

[123] YU S H, ZHANG H G, MING Z Y, et al. Optimal control for continuous-time unknown nonlinear affine systems: a Q-learning approach[J]. IEEE Transactions on Automation Science and Engineering, 2024, 21(4): 6519-6527.

[124] LIU Y Y, WANG Z S. Reinforcement learning-based tracking control for a class of discrete-time systems with actuator fault[J]. IEEE Transactions on Circuits and Systems II: Express Briefs, 2022, 69(6): 2827-2831.

[125] LIU D R, WEI Q L, WANG D, et al. Adaptive dynamic programming with applications in optimal control[M]. Cham, Switzerland: Springer International Publishing, 2017.

[126] PARIKH A, KAMALAPURKAR R, DIXON W E. Integral concurrent learning: Adaptive control with parameter convergence using finite excitation[J]. International Journal of Adaptive Control and Signal Processing, 2019, 33(12): 1775-1787.

[127] AWAN A U, ZAMANI M. Formal synthesis of safety controllers for unknown systems using Gaussian process transfer learning[J]. IEEE Control Systems Letters, 2023, 7: 3741-3746.

[128] CHEN Z, LIANG X, ZHENG M H. Knowledge transfer between different UAVs for trajectory tracking[J]. IEEE Robotics and Automation Letters, 2020, 5(3): 4939-4946.

[129] ZHANG H R, ZHAO C H. Stable transfer learning-based control: an off-dynamics adaptive approach for unknown nonlinear systems[J]. Neurocomputing, 2025, 616: 128951.

[130] O'CONNELL M, CHO J, ANDERSON M, et al. Learning-based minimally-sensed fault-tolerant adaptive flight control[J]. IEEE Robotics and Automation Letters, 2024, 9(6): 5198-5205.

[131] MCCLEMENT D G, LAWRENCE N P, BACKSTRÖM J U, et al. Meta-reinforcement learning for the tuning of PI controllers: an offline approach[J]. Journal of Process Control, 2022, 118: 139-152.

[132] WANG D, HE H B, LIU D R. Adaptive critic nonlinear robust control: a survey[J]. IEEE Transactions on Cybernetics, 2017, 47(10): 3429-3451.

[133] 范泉涌. 基于神经网络的非线性系统自适应容错控制方法研究[D]. 沈阳: 东北大学, 2017.
FAN Q Y. Research on adaptive fault-tolerant control for nonlinear systems based on neural networks[D]. Shenyang: Northeastern University, 2017. (in Chinese)

[134] HORNG J H. Neural adaptive tracking control of a DC motor[J]. Information Sciences, 1999, 118(1/2/3/4): 1-13.

[135] PARK J H, HUH S H, KIM S H, et al. Direct adaptive controller for nonaffine nonlinear systems using self-structuring neural networks[J]. IEEE Transactions on Neural Networks, 2005, 16(2): 414-422.

[136] KINGRAVI H A, CHOWDHARY G, VELA P A, et al. A reproducing Kernel Hilbert Space approach for the

online update of Radial Bases in neuro-adaptive control[C]//Proceedings of the 50th IEEE Conference on Decision and Control and European Control Conference, 2011: 1796-1802.

[137] YANG Y L, ZHU H F, ZHANG Q C, et al. Sparse online kernelized actor-critic Learning in reproducing kernel Hilbert space[J]. Artificial Intelligence Review, 2022, 55(1): 23-58.

[138] MENACHE I, MANNOR S, SHIMKIN N. Basis function adaptation in temporal difference reinforcement learning[J]. Annals of Operations Research, 2005, 134(1): 215-238.

[139] BRUNKE L, GREEFF M, HALL A W, et al. Safe learning in robotics: from learning-based control to safe reinforcement learning[J]. Annual Review of Control, Robotics, and Autonomous Systems, 2022, 5: 411-444.

[140] GARC ÍA J, FERNÁNDEZ F. A comprehensive survey on safe reinforcement learning[J]. Journal of Machine Learning Research, 2020, 16: 1437-1480.

[141] MARVI Z, KIUMARSI B. Safe reinforcement learning: a control barrier function optimization approach[J]. International Journal of Robust and Nonlinear Control, 2021, 31(6): 1923-1940.

[142] MATIISEN T, OLIVER A, COHEN T, et al. Teacher-student curriculum learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(9): 3732-3740.

[143] STONE P, SUTTON R S, KUHMUENCH G, et al. Curriculum learning for reinforcement learning domains: a framework and survey[J]. Journal of Machine Learning Research, 2021, 21(181): 1-50.

[144] VAN WESEL P, GOODLOE A E. Challenges in the verification of reinforcement learning algorithms[R]. Hampton, USA: NASA Langley Research Center, 2017.

[145] SHEN Y X, WANG Z D, DONG H L, et al. Joint state and unknown input estimation for a class of artificial neural networks with sensor resolution: an encoding–decoding mechanism[J]. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(2): 3671-3681.

[146] CHEN J H, ZHAO C H, SONG P Y, et al. Unified low-dimensional subspace analysis of continuous and binary variables for industrial process monitoring[J]. IEEE Transactions on Cybernetics, 2025, 55(3): 1135-1146.

[147] CHEN C, XIE L H, XIE K, et al. Adaptive optimal output tracking of continuous-time systems via output-feedback-based reinforcement learning[J]. Automatica, 2022, 146: 110581.

[148] ÅSTRÖM K J. Optimal control of Markov processes with incomplete state information[J]. Journal of Mathematical Analysis and Applications, 1965, 10(1): 174-205.

[149] WHITE C C. A survey of solution techniques for the partially observed Markov decision process[J]. Annals of Operations Research, 1991, 32(1): 215-230.

[150] MODARES H, LEWIS F L, JIANG Z P. Optimal output-feedback control of unknown continuous-time linear systems using off-policy reinforcement learning[J]. IEEE Transactions on Cybernetics, 2016, 46(11): 2401-2410.

[151] NI T W, MA M, EYSENBACH B, et al. When do transformers shine in RL? decoupling memory from credit assignment[C]//Proceedings of the 37th International Conference on Neural Information Processing Systems, 2023.

[152] CHEN L L, LU K, RAJESWARAN A, et al. Decision transformer: reinforcement learning via sequence modeling[C]//Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021), 2021.

[153] JANNER M, LI Q, LEVINE S. Offline reinforcement learning as one big sequence modeling problem[C]//Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021), 2021.

[154] AMES A D, XU X R, GRIZZLE J W, et al. Control barrier function based quadratic programs for safety critical systems[J]. IEEE Transactions on Automatic Control, 2017, 62(8): 3861-3876.

[155] AZIMI V, HUTCHINSON S. Exponential control Lyapunov-barrier function using a filtering-based concurrent learning adaptive approach[J]. IEEE Transactions on Automatic Control, 2022, 67(10): 5376-5383.

[156] AMES A D, COOGAN S, EGERSTEDT M, et al. Control barrier functions: theory and applications[C]//Proceedings of the 18th European Control Conference (ECC), 2019: 3420-3431.

[157] FAN Q Y, YANG G H. Adaptive nearly optimal control for a class of continuous-time nonaffine nonlinear

systems with inequality constraints[J]. ISA Transactions, 2017, 66: 122-133.

[158] JACOBSON D, LELE M. A transformation technique for optimal control problems with a state variable inequality constraint[J]. IEEE Transactions on Automatic Control, 1969, 14(5): 457-464.

[159] ZHANG H R, ZHAO C H, DING J L. Robust safe reinforcement learning control of unknown continuous-time nonlinear systems with state constraints and disturbances[J]. Journal of Process Control, 2023, 128: 103028.

[160] BIAN T, JIANG Z P. Value iteration and adaptive dynamic programming for data-driven adaptive optimal control design[J]. Automatica, 2016, 71: 348-360.

[161] XIE K D, ZHENG Y W, LAN W Y, et al. Adaptive optimal output regulation of unknown linear continuous-time systems by dynamic output feedback and value iteration[J]. Control Engineering Practice, 2023, 141: 105675.

[162] ANNASWAMY A M, GUHA A, CUI Y N, et al. Integration of adaptive control and reinforcement learning for real-time control and learning[J]. IEEE Transactions on Automatic Control, 2023, 68(12): 7740-7755.

[163] YUKSEK B, INALHAN G. Reinforcement learning based closed-loop reference model adaptive flight control system design[J]. International Journal of Adaptive Control and Signal Processing, 2021, 35(3): 420-440.

[164] ZHANG J F, ZHANG H R, ZHAO C H. Toward universal controller: performance-aware self-optimizing reinforcement learning for discrete-time systems with uncontrollable factors[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2025, 55(5): 3249-3260.

[165] ZHANG J F, ZHAO C H, DING J L. Deep reinforcement learning with domain randomization for overhead crane control with payload mass variations[J]. Control Engineering Practice, 2023, 141: 105689.

[166] HELWA M K, SCHOELLIG A P. Multi-robot transfer learning: a dynamical system perspective[C]//Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2017: 4702-4708.