DOI：10.13334/j.0258-8013.pcsee.223433 文章编号：0258-8013 (2024) 01-0001-21 中图分类号：TM 73；TP 181 文献标识码：A

电力系统优化控制中强化学习方法应用及挑战

毕聪博[1]，唐聿劼[2]，罗永红[1]，陆超[1*]

(1. 新型电力系统运行与控制全国重点实验室(清华大学电机工程与应用电子技术系)，北京市 海淀区 100084；2. 北京大学工学院工业工程与管理系，北京市 海淀区 100871)

Review on Critical Problems in Reinforcement Learning Methods Applied in Power System Optimization and Control Scenarios

BI Congbo[1], TANG Yujie[2], LUO Yonghong[1], LU Chao[1*]

(1. National Key Laboratory of New-type Power System Operation and Control (Department of Electrical Engineering, Tsinghua University), Haidian District, Beijing 100084, China; 2. Department of Industrial Engineering & Management, Peking University, Haidian District, Beijing 100871, China)

ABSTRACT: Reinforcement learning (RL) method has been applied in some fields of power system. The applications in power system optimization and control scenarios show admirable results. However, there are still some critical problems in the process of applying reinforcement learning methods to real-world power system applications. This paper first summarizes the basic theory and the state-of-art progress of reinforcement learning. Then, some critical problems in applications of reinforcement learning in various optimization and control scenarios in power system are pointed out. Finally, some future directions of reinforcement learning applied to power system decision-making and control scenarios are discussed.

KEY WORDS: reinforcement learning(RL); power system; optimization and control

摘要：强化学习(reinforcement learning，RL)方法目前已应用于电力系统的多个领域，在电力系统优化与控制领域的一些应用展现出良好的结果。但在强化学习方法落地于实际电力系统应用的过程中依然存在一些关键性问题。该文首先概述强化学习基础理论与研究现状，随后提出强化学习理论落地于电力系统各领域优化与控制过程中存在的关键问题。最后探讨强化学习应用于电力系统优化与控制的研究展望。

关键词：强化学习(RL)；电力系统；优化与控制

0 引言

我国新一代电力系统的主要特征包括高比例可再生能源接入、高比例电力电子设备接入、支撑多能互补的综合能源网以及电力系统与信息通信技术深度融合[1]。高比例可再生能源出力的随机性、间歇性和波动性使系统运行方式呈现多样性和多变性[2]，电力电子设备快速响应引起的宽频振荡严重威胁电力系统运行安全[3]，电力系统的高度信息化则对海量数据的高效处理提出了更高的要求[4]。在此背景下，新型电力系统的结构和运行特性将变得日益错综复杂，其控制与优化面临着新的严峻挑战[5]。受模型简化等条件制约，传统基于模型的机理性方法难以有效应对实际系统复杂多变的运行场景，亟需更高效、更灵活、更安全的电力系统优化与控制方法。

数据驱动的强化学习(reinforcement learning，RL)方法并不依赖于具体的模型参数，而是通过类似人类学习的过程，在与环境的在线交互逐渐学习关于环境的“知识”，并不断修正智能体对自身策略的评价及更新优化对应的策略[6-7]。由于具有在线交互性好、在高度复杂的控制决策场景中表现优秀、与最优控制等学科融合紧密等特点，强化学习在机器学习领域具有重要的地位[8-10]。近年来，随着大数据技术的进步和计算能力的提升，融合强化学习和深度学习两者优势的深度强化学习方法(deep reinforcement learning，DRL)发展迅速，目前已经应用于高性能游戏[11-13]、蛋白质折叠预测[14]、核聚变控制[15]等多个复杂场景的决策控制。强化学习应用于电力系统优化与控制的研究已经在多个领域和层次展开[16-26]，涵盖了频率控制、电压稳定

基金项目：国家自然科学基金项目(U2066601，52242701)。
Project Supported by National Natural Science Foundation of China (U2066601, 52242701).

控制、电力系统运行优化、电力市场、综合能源系统管理、电力系统信息安全、需求响应等多个方面。事实上，强化学习与现代控制领域中的近似动态规划方法(approximate dynamic programming，ADP)有着紧密的联系[27]，后者已经在电力系统的直流控制、调频控制等场景中获得了广泛的应用[28-31]。在某些感知度较低、随机性较强或机理研究仍未成熟的应用场景[32-36]中，强化学习方法依然展现了其强大的适应能力。以低压减载为例，相比于传统基于操作规程的固定的负荷切除量，强化学习可以实现更精确的负荷切除量以减小负荷损失，同时在线运行决策的耗时接近于传统方法[37-38]；而相对于传统的人工设定运行水平以及预想故障集的控制策略，强化学习方法可以对环境特性进行更深入的挖掘，且具有更好的对未知故障的适应性[34,39-40]。

但是，强化学习在应用于现实的物理系统的过程中依然存在诸多问题。Gabriel[41]指出了强化学习应用于现实世界系统的九大挑战，包括从有限的样本中学习实际系统、与部分可观测的系统进行交互等。虽然近年来强化学习在电力系统多个领域的优化与控制场景中表现出良好的性能，但在落地于实际电力系统应用的过程中，强化学习依然面对一些比较关键的问题和挑战，如某些电力系统优化控制的应用场景不一定严格满足马尔科夫属性、强化学习应用于部分可观系统的可行性尚待明晰、缺乏有关强化学习方法应用于某些优化控制场景的最优性或次优性保证，以及如何应对仿真环境和实际电力系统存在的模型偏差和样本偏差、如何增强强化学习方法的可解释性等。对上述问题的研究有助于推动强化学习方法在实际电力系统中的落地应用。

针对上述问题，本文结合强化学习的理论基础和研究现状，总结强化学习应用于电力系统优化与控制的一些关键问题与挑战，并针对性地提出研究展望。文章的其余部分结构如下：1 节介绍强化学习的基础理论及研究现状；2 节探讨强化学习应用于电力系统优化与控制的问题与挑战，并分别分析每个问题的重要性以及对应问题在电力系统不同领域优化与控制应用中的具体表现，为后续研究提供借鉴；3 节在前述问题分析基础上，对强化学习应用于电力系统优化控制的未来研究方向提出展望；最后是总结。

1 强化学习基础理论

强化学习是一类无需具体环境模型参数的、通过与环境交互获取知识的自监督的机器学习方法。典型的强化学习框架如图 1 所示，包括环境和智能体两大部分。环境通常被建模为马尔科夫决策过程(Markov decision process，MDP)，智能体获取当前时刻的环境观测值并依据历史经验做出决策，通过控制动作使环境发生变化。实际上，可以将智能体看作由感知模块、策略模块、价值函数模块和转换模型 4 个部分组成的有机整体[42]。

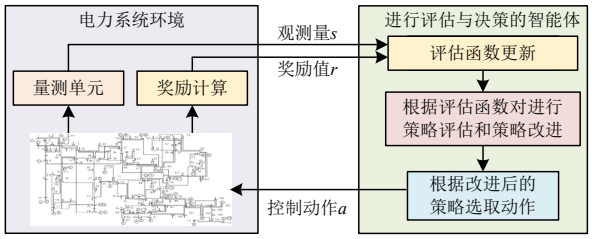


图 1 强化学习方法框架

Fig. 1 Framework of reinforcement learning method

1.1 马尔可夫性与马尔科夫决策过程

大部分强化学习方法要求交互的环境具有马尔科夫性。当实随机过程 $X(t)$，$t\in T$ 满足：

$$P\{X(t_{n+1})\,|\,X(t_n),...,X(t_1)\}=P\{X(t_{n+1})\,|\,X(t_n)\} \quad (1)$$

即在该随机过程中，下一时刻的状态 $X(t_{n+1})$仅与当前时刻状态 $X(t_n)$有关，而与之前的状态 $X(t_{n-1})$, $X(t_{n-2})$, …, $X(t_1)$无关，则称该随机过程具有马尔科夫性[43]。

MDP 以马尔科夫性为基础。包含奖励的有限 MDP 以 5 元组$\{S,A,P,R,\gamma\}$描述[43]。一个典型的 MDP 如图 2 所示。其中，状态空间 S 为该过程的状态集合，时刻 t 的状态表示为 s_t, $s_t\in S$。动作空间 A 是决策过程中的动作集合，时刻 t 的状态表示为 a_t, $a_t\in A$。$P(s,s',a)=P(s_{t+1}=s'|s_t=s,a_t=a)$描述了环境的动态特性，表示当前状态为 s 时执行动作 a 后转移到状态为 s' 的概率，不同的状态和动作组合(s_t,a_t)对应于不同的

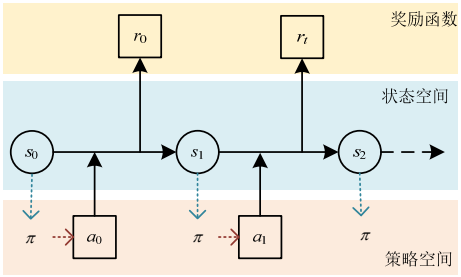


图 2 马尔科夫决策过程示意

Fig. 2 Illustration of a Markov decision process

状态转移。$R(s,a,s')$表示在动作 a 下从状态 s 转移到状态 s'对应的奖励，t 时刻的奖励记作 $r_t=R(s_t,a_t,s_{t+1})$。策略π是从状态空间 S 到动作空间 A 上的概率分布，分为确定性策略和随机性策略，前者可写作$\pi(s)=a$，后者可写作$\pi(a|s)=P(a_t=a|s_t=s)$。智能体的目标是获取尽可能高的长期收益，对应地需要平衡远期收益与当前奖励的影响。一种方法是引入折扣因子 $\gamma\in[0,1)$，如果γ越小，表明智能体更重视短期奖励，反之则更重视累计收益。强化学习算法的优化目标是从 0 时刻开始的期望累计收益 $\mathbb{E}(G)$：

$$\mathbb{E}(G)\doteq\mathop{\mathbb{E}}_{\tau\sim\pi_\theta}\left(\sum_{t=1}^{\infty}\gamma^t R_t\right) \tag{2}$$

1.2　基于值函数和基于策略函数的强化学习算法

根据智能体是否从数据构建环境模型，强化学习方法分为基于模型和不基于模型的两大类方法。考虑到强化学习应用于电力系统的背景之一是日益复杂的运行场景和电力系统模型，本文主要讨论不基于模型的强化学习方法在电力系统中的应用。不基于模型的强化学习方法又可以分为基于值函数和基于策略的两大类算法[44]。

基于值函数的强化学习方法直接学习最优的 Q 函数，而基于 Q 函数的贪心策略即为当前评估下的最优策略。环境的状态空间和动作空间都有限的情况下，可以采用表格形式的 Q 函数。典型的基于值函数的表格型强化学习算法包括同轨策略的 SARSA (state action reward state action)算法[45]和离轨策略的 Q 学习算法(Q-learning)等[46]。

针对复杂高维环境，以含参数 θ 的函数 $Q_\theta(s,a)$ 替代 Q 表格，可以缓解维数爆炸的问题。以深度神经网络作为 Q 函数，即深度 Q 网络(deep Q-network, DQN)算法[47-48]，结合神经网络的数据挖掘能力，可以实现复杂情境下的高效决策。

相较于基于值函数的强化学习方法对值函数进行优化的做法，基于策略的强化学习方法直接对参数化的策略 π_θ 进行改进。基于策略的强化学习以策略梯度(policy gradient, PG)算法为基础。文献[49]提出参数化的策略 π_θ 对应的梯度可由采样估计。基于策略梯度理论的 REINFORCE 算法[50]是经典的强化学习算法之一。

结合策略梯度与演员−评论家(actor-critic，AC)框架的 AC 算法[51]引入了评价 Actor 行动效果的 Critic，Actor 基于状态计算对应的动作，Critic 估计动作的 Q 值或 V 值。在 AC 算法的策略迭代中引入约束，得到信赖域策略优化(trust region policy optimization，TRPO)算法与近端策略优化(proximal policy optimization，PPO)算法等。

前述的策略梯度方法均基于随机策略，即输出动作的概率分布。在动作空间连续的情况下，无法通过枚举计算各动作对应的 Q 函数与选择最优动作。确定性策略可以直接基于状态 s 输出确定的动作 $a=\pi(s)$，无需进行最优动作的选择，可应用于动作空间连续的环境。以深度确定性策略梯度(deep deterministic policy gradient，DDPG)算法[52]为例，其利用卷积神经网络近似策略函数和 Q 函数，其中策略网络输出确定的动作。

1.3　结合函数近似的强化学习及理论风险

随着强化学习应用场景日益复杂，状态空间和动作空间的维度增加甚至变为连续，表格型强化学习方法面临计算量急剧增加的“维数灾”等问题[53]。结合函数近似结构的强化学习方法获得了学术界的关注[54]。强化学习中的函数近似主要分为对值函数近似、对策略函数近似和对两者同时近似 3 种类型。值函数近似方面，Parr[55]提出了线性价值函数逼近等价于线性模型逼近的一种形式，在一定意义上说明了函数近似与强化学习方法结合的合理性；Melo[56]提出了 Q 函数的一些变体与函数近似结合时的收敛特性。策略函数近似方面，Sutton[49]提出了函数近似的策略梯度，当 Q 函数的近似函数 $f_\omega(s,a)$ 满足一定的条件时，可以用 $f_\omega(s,a)$ 替换原策略梯度中的 $Q_\pi(s,a)$。Agarwal[57]则给出了策略函数近似情况下应用策略梯度方法的收敛速度和误差分析。

为解决高维决策问题，融合深度学习感知能力与强化学习决策能力的深度强化学习方法近年来成为应用主流[58]。深度强化学习结合了 2 者的长处，在复杂的控制优化场景中相较于传统的优化控制方法在求解时间上具有优势，而相较于现有的人工制定的操作规程又更加精准，在某些场景中甚至表现要远高于人类[47-48]。但函数近似的引入在缓解了维数灾难、提高应用灵活性的同时，也带来了非平稳性、发散风险、最优性和泛化性能缺乏理论保证等新的问题和挑战[59-60]。特别地，深度强化学习方法常常会面临同时使用函数近似、异策略和自举 3 种方法带来的发散风险[61]。研究深度强化学习算法的收敛性[62–64]，以及如何构造合理的近似函数结构以保证训练过程的安全性[65–68]成为理论界研究的热点问题之一。

2 强化学习应用于电力系统优化与控制的问题与挑战

强化学习应用于电力系统优化与控制的相关研究已经在多个领域展开，但在强化学习理论落地于电力系统优化与控制的过程中还存在一些问题与挑战。基函数选取困难、学习速度慢及算法收敛性难以保证是强化学习应用中的关键问题，也是强化学习自身存在的理论性缺陷。除上述关键问题以外，强化学习落地于电力系统优化与控制还存在一系列重要问题。强化学习应用于电力系统优化控制的前提是应用场景具有马尔科夫性质，目的是在高维复杂场景中实现有一定理论性能保证的优化控制。同时，作为实际的物理系统，电力系统的某些优化控制场景是部分可观测的，部分可观性对场景建模与最优解求解均会有一定影响。因此，环境是否满足马尔科夫属性、是否有应用于部分可观场景的可行性、学习过程是否有最优性或次优性保证是强化学习落地于电力系统优化与控制的重要问题。本节将总结强化学习应用于电力系统优化与控制的 3 个重要问题及其研究现状，同时简要概述其他影响强化学习落地于电力系统优化与控制场景的问题。

2.1 电力系统优化与控制过程的马尔科夫属性问题

主流强化学习方法的应用前提之一是与智能体交互的环境可以建模为 MDP。如果与智能体交互的电力系统应用场景不具有马尔科夫性，则价值函数的迭代计算无法成立，后续强化学习方法应用于对应场景的理论基础更无从谈起。

电力系统是高维、复杂、随机性强的物理系统，不同的应用场景中存在大量的物理过程。某些优化与控制过程并不严格满足马尔科夫属性，但是经过一定的简化之后也可以建模成为 MDP。本节针对不同的应用场景对其状态转移与马尔科夫属性进行分析与综述，表 1 列举了强化学习应用于电力系统不同领域优化与控制过程中状态转移与马尔科夫属性情况的文献摘要[69-92]。

表 1 电力系统优化与控制过程 Markov 属性的文献摘要

Table 1 Literature summary on Markov properties of power system optimization and control scenarios

文献	研究领域	研究场景	状态转移数学形式	状态转移有不确定因素	马尔科夫属性
文献[69]	电压控制	分布式电源就地自适应电压控制	代数方程	否	满足
文献[70]	电压控制	一级自动电压控制	微分方程	否	满足
文献[71]	电压控制	拓扑变化的电力系统电压稳定控制	离散化的微分方程	否	满足
文献[72]	频率控制	风电系统自适应负荷频率控制	微分方程	否	满足
文献[73]	频率控制	多区域负荷频率控制	微分方程组	否	满足
文献[74]	频率控制	紧急频率控制	微分方程	否	满足
文献[75]	优化调度	配电网无功优化	代数方程	否	满足
文献[76]	优化调度	无功-电压协调控制	代数方程	否	满足
文献[77]	优化调度	电网断面极限传输能力趋优控制	微分代数混合方程	是	简化后满足
文献[78]	综合能源系统管理	可再生能源的经济调度	代数方程	是	简化后满足
文献[79]	综合能源系统管理	综合能源系统动态经济调度	代数方程	是	简化后满足
文献[80]	综合能源系统管理	智能微电网群控制优化	代数方程	是	简化后满足
文献[81]	综合能源系统管理	储能系统优化运行	代数方程	是	简化后满足
文献[82]	电力市场	基于需求响应的电力市场动态定价	代数方程	是	简化后满足
文献[83]	电力市场	电力现货市场定价机制	代数方程	是	简化后满足
文献[84]	信息安全	电网故障序列搜索及防御	代数方程	是	简化后满足
文献[85]	信息安全	网络攻击下对分布式能源的调度	代数方程	是	简化后满足
文献[86]	信息安全	薄弱线路辨识	代数方程	是	简化后满足
文献[87]	其他	潮流计算条件调整	代数方程	否	满足
文献[88]	其他	静态电压稳定裕度评估	代数方程	否	满足
文献[89]	其他	电力系统静态安全预防控制	代数方程	否	满足
文献[90]	其他	输电网络扩展规划	代数方程	是	简化后满足
文献[91]	其他	超短期光伏功率预测	代数方程	是	简化后满足
文献[92]	其他	负荷参数校准	微分方程	是	简化后满足

2.1.1　状态转移确定的应用场景

电力系统中的一部分优化与控制过程可由微分代数方程组(differential algebraic equations，DAEs)描述[93]，假定 DAEs 中涉及的状态变量均为可观测的，根据当前时刻的状态 $x(t)$，以及对 DAEs 进行离散化得到的描述状态转移的差分代数方程组，求解得到后继时刻的状态量 $x(t+\Delta t)$，仅与当前状态与动作有关。由于时变负荷变化缓慢，因此认为研究过程中的负荷不变，上述物理过程满足马尔科夫性。此外，电力系统的某些长时间尺度的优化调度过程可建模为由代数方程描述状态转移的序列决策过程，也具有马尔科夫性。

1）电压控制。根据研究时间尺度不同，电压控制问题的状态转移过程以代数方程组或 DAEs 描述。在文献[69]中，作为状态特征量的有功、无功功率与母线电压满足潮流方程约束，后继状态可根据当前时间断面的系统运行状态以及当前时刻的调节确定；文献[70]将发电机的励磁电压控制过程建模为时域离散的差分方程，可根据状态特征量和动作计算下一时刻状态特征量；文献[71]研究拓扑变化的电压稳定控制问题，在给定的故障场景下，电力系统后继时刻的电压可由当前的运行状态和切除的负荷量 2 者共同决定。

2）频率控制。现有的频率控制方法包括发电机本地频率控制、自动发电控制和经济调度等，前 2 者的状态转移过程以 DAEs 描述[94]。文献[72]以时域离散化的微分方程描述风电系统自适应负荷频率控制过程的状态转移；文献[73]将负荷频率控制过程建模为包含控制动作的离散化非线性微分方程，作为状态量的区域控制误差的状态转移由上述的微分方程确定；文献[74]中电力系统被建模成多输入多输出高阶非线性系统，对应的状态转移过程由系统动态微分方程组唯一确定。

3）优化调度。电力系统中的部分优化调度问题的状态转移由代数方程描述，具有完备的马尔科夫属性。如文献[75]将配电网的持续无功优化问题转变为一个多步 MDP，状态转移过程由潮流方程描述，后继运行状态仅与上一时刻的运行状态与控制动作有关；文献[76]中，三相不平衡配电网的电压-无功综合控制的状态转移过程由配电网潮流方程描述，后继状态可以直接由当前运行状态和控制动作计算。但部分涉及不确定性的优化调度场景需要对不确定因素进行简化才可以应用 MDP 建模，如文献[77]基于历史数据对输电线路负荷的不确定性进行建模。

4）其他应用。文献[87]利用强化学习方法调整大电网潮流计算的初始条件，潮流计算条件调整问题被建模为 MDP，状态空间为潮流计算的初始条件与区域特性，其状态转移由边界条件的调整动作确定；文献[88]提出了一种基于强化学习理论的静态电压稳定裕度评估方法，将最优潮流计算过程建模为逼近稳定边界的多步序列决策；文献[89]则将电力系统静态安全预防控制问题表述为序列决策问题，状态转移过程由电力系统潮流方程组确定。

2.1.2　状态转移具有较强不确定性的应用场景

电力系统中的一部分优化控制问题过程建模为马尔科夫决策过程的过程中存在状态转移概率矩阵是时变的、后继状态与当前状态-动作组合无关或者弱相关等问题，即状态转移具有较强的不确定性，并不严格满足马尔科夫性。对不确定成分进行简化可以将场景建模为 MDP，从而应用强化学习方法。

1）综合能源系统(integrated energy system，IES)管理。大部分 IES 的调度控制具有较强的不确定性，MDP 建模过程中一般将随机因素转化为服从一定分布的随机变量。文献[78]基于历史数据将随机的间歇性电源等效为满足一定分布的随机性电源，同时在 MDP 建模中加入时序相关因素以减小对不确定性的预测误差；文献[79]忽略负荷需求和光伏发电的随机性与负荷的时变性，将不确定的用户电热负荷需求和预测的光伏发电简化为确定且已知的观测量，将时变的马尔科夫转移简化为一定概率的状态转移；文献[80]利用神经网络在线估计未知内部阻抗的混合储能系统的动态特性，另一个神经网络通过在线学习，根据估计的系统动态，计算出混合储能系统的最佳控制输入；文献[81]对光储充电站储能系统优化运行进行研究，将储能系统的可用容量衰减过程简化为基于历史统计数据计算的随机过程。

2）电力市场。电力市场中的定价问题或需求侧响应问题通常被建模成序列决策问题或多智能体博弈问题，其中负荷侧的需求或发电侧的供给中存在的不确定性常常被简化为一定的概率分布。文献[82]中，动态定价问题被表述为离散有限马尔可夫决策过程，其中假定奖励和能源消耗仅取决于相

应时间段的预测的能源需求和零售价格，与历史数据无关；文献[83]中则利用蒙特卡洛方法对微电网中存在的不确定性进行估计，能源零售商根据估计结果选择其零售定价策略。

3）数据攻击与防御。针对电力系统数据安全性的网络攻击过程通常被简化建模为有限状态空间与动作空间的序列决策过程。文献[84]将电网故障序列搜索及防御过程建模为多阶段动态零和博弈模型，其中假定连锁故障具有马尔科夫特性；文献[85]研究网络攻击下对分布式能源的调度策略，其中攻击者的行为被简化为离散时间序列下的攻击动作，针对性的控制策略也被建模为单一代理的离散时间序列下的离散动作；文献[86]则基于双 Q 学习方法，利用序列数据攻击对电网薄弱线路进行辨识。上述工作的共同点是研究分钟级时间尺度的攻击过程，忽略了暂态过程和运行水平变化对电网安全性的影响，同时限定攻击者一次只能对 1 条线路发动攻击。简化后的模型具有马尔科夫属性。

4）其他应用。文献[90]提出基于强化学习的输电网络规划方案，其中负荷需求和故障被简化为满足一定概率分布的随机量；文献[91]利用强化学习和组合式深度学习模型进行超短期光伏功率预测，输入的状态量为不同模型对下一阶段的预测量，其中假定计算得到的优化加权权重变化与当前的权重满足一定的概率分布；文献[92]基于强化学习方法对模型负荷参数进行辨识，第一阶段对不确定性较强的模型负荷参数空间进行离散化以确定接近真实瞬态动态的适当负载组合，第 2 阶段再利用蒙特卡洛模拟选择负荷模型其余的参数值。

2.1.3 小结

综上，以确定的代数方程组或 DAEs 描述状态转移的电力系统优化或控制过程可根据当前状态量和动作基于状态方程计算后继时刻的状态量，满足马尔科夫属性。一部分优化或控制场景由于具有存在状态转移概率矩阵是时变的、后继状态与当前状态-动作组合无关或者弱相关等因素，无法直接应用 MDP 进行建模，一般基于历史信息或直接将不确定因素简化为满足一定分布的随机变量再进行建模，简化后的模型具有马尔科夫性。在不确定因素对研究场景没有本质影响的前提下，简化不确定因素的影响，将环境简化建模为 MDP 并应用强化学习方法是合理可行的。

2.2 强化学习应用于电力系统优化与控制场景中的可观性问题

随着传感与量测系统的发展及电力系统仿真技术的进步，大量实测数据和仿真数据为以深度强化学习为代表的数据驱动方法应用于电力系统提供了坚实的数据基础。但实际电力系统的某些物理过程的状态转移方程中状态空间维数很大，某些情况下即使状态转移过程本身满足马尔科夫属性，单一时间断面的观测数据是否能完整反映系统的运行状态也有待商榷。如果观测空间 O 和状态空间 S 并不等价，则无法直接根据观测量推断准确的隐藏状态，进而导致无法直接计算对应的状态价值函数，后续的基于强化学习方法的策略评估与优化将更无从谈起。

依据历史时间序列构筑信念状态(belief state，BS)[95]或利用数据挖掘方法进行时序特征提取[96]，是处理部分可观性常用的手段。经典的两种针对部分可观测的建模方法是将环境建模为部分可观马尔科夫决策过程(partially observable markov decision process，POMDP)[97]和预测状态表示(predictive state representations，PSR)[98]。如何有效利用历史信息求解 POMDP 最优策略已成为强化学习领域的热点问题之一[99–101]，类似的处理方法在电力系统中也有应用[102]。

实际电力系统中的很多场景常常不能实现理想的完全可观测，因此确定电力系统优化与控制场景的可观性，以及研究部分可观场景下直接基于量测数据驱动的强化学习方法依然具有重要的实际意义。表 2 列举了电力系统优化与控制场景中的可观性的文献摘要[34-35,102-115]。

2.2.1 完全可观测的应用场景

电力系统中的一部分优化与控制的应用场景，尤其是一些时间尺度较长的、以代数方程描述状态转移的应用场景，大多具有完全可观测性。典型的如静态电压控制、一部分系统优化调度和一部分综合能源系统管理场景等。

静态电压控制方面，文献[103]将较慢时间尺度的电压控制过程建模为 MDP 过程，根据观测量即节点负荷与上一时刻的电容器配置，可以唯一确定系统运行状态；在文献[104]中，定义系统母线电压幅值、相角，线路有功、无功功率，发电机的出力，母线负荷为观测量，在潮流方程约束下系统运行状态被唯一确定。

表 2　强化学习应用于电力系统优化与控制场景中的可观性的文献摘要

Table 2　Literature summary on observability of reinforcement learning methods applied to power system optimization and control scenarios

文献	研究领域	研究场景	问题可观测性	问题建模形式/部分可观解决方案
文献[103]	静态电压控制	电力系统两阶段电压控制	完全可观测	MDP/—
文献[104]	静态电压控制	配电网变压器有载分接头控制	完全可观测	MDP/—
文献[35]	优化调度	配电网无功优化	完全可观测	MDP/—
文献[105]	优化调度	电网运行与维护	完全可观测	MDP/—
文献[106]	综合能源系统管理	风电场储能系统预测决策调度	完全可观测	MDP/—
文献[107]	综合能源系统管理	微电网复合储能协调控制	完全可观测	MDP/—
文献[108]	暂态电压控制	考虑暂态电压稳定性的无功补偿	部分可观测	POMDP/时序状态扩充
文献[34]	暂态电压控制	防止系统电压失稳的低压减载	部分可观测	POMDP/ConvLSTM
文献[109]	频率控制	多微网负荷频率协同控制	部分可观测	POMDP/ MARL
文献[102]	频率控制	实时闭环广域分散电力系统稳定器的设计	部分可观测	POMDP/时序状态扩充+MARL
文献[110]	紧急控制	发电机跳闸控制	部分可观测	POMDP/时序状态扩充
文献[111]	紧急控制	故障引起的延迟电压恢复	部分可观测	POMDP/LSTM
文献[112]	综合能源系统管理	微电网中的分布式能源管理	部分可观测	POMDP/MARL
文献[113]	综合能源系统管理	分布式电源协同优化	部分可观测	POMDP/MARL
文献[114]	信息安全	数据完整性攻击的防御	部分可观测	POMDP/时序数据扩充
文献[115]	信息安全	针对智能电网的网络攻击在线检测	部分可观测	POMDP/时序数据扩充

优化调度方面，文献[35]定义观测空间包括配电系统内各母线电压、各调节设备投切档位和各调节设备已经完成的动作，观测空间完整反映了系统的运行状态；文献[105]研究具有预测和健康管理能力的电网的运行和维护，观测空间被定义为电力系统组件是否退运或其运行水平，可以据此确定系统运行状态。

综合能源系统管理方面，文献[106]以前瞻电价、储能系统存储电量和风电场的量测数据作为风电场储能系统调度过程的观测变量，可以完整描述风电场储能系统的运行状态及预测后继状态；文献[107]研究强化学习应用于微电网复合储能协调控制，选取的状态空间可以唯一确定微电网的运行水平。

2.2.2　部分可观测的应用场景

电力系统中的很多过程是由微分-代数方程描述的动态过程，计算某一时刻的后继状态需要观测到该时刻断面完整的状态变量信息，但上述观测在实际电力系统中较难实现，通常此类运行场景具有部分可观测性质，例如暂态电压控制和紧急控制场景等。此外，一部分场景中智能体无法实现对环境信息的完整感知，例如分布式控制和针对电网的安全攻击等。应对部分可观场景的常用方法包括引入历史信息作为增广状态的一部分、放弃理论最优性保证、将问题建模为 POMDP 等；而多智能体可以通过通信实现信息共享，一定程度上缓解了单智能体观测范围有限的弊端。

暂态电压控制方面，文献[108]研究考虑暂态电压稳定性的无功补偿方案，考虑到单一时间断面的观测无法完整反映系统的动态特性，将一段时间 T 内的节点电压幅值、频率、有功和无功功率作为智能体的信息输入；文献[34]将 DRL 方法应用于低压减载问题，输入信息为电力系统的节点电压幅值、有功和无功功率，传输线功率以及反映系统电压变化特性的雅可比矩阵，在上述观测变量无法表征系统动态的情况下采用卷积长短时记忆(convolutional long short-term memory，ConvLSTM)网络处理时空信息。频率控制方面，文献[109]将自身的实时频率偏差、柔性负荷的充电功率上下限以及其余智能体的动作集合作为单个智能体的观测空间，利用 MADDPG 方法弥补单个智能体只能实现部分观测的缺陷；文献[102]则研究电力系统稳定器的设计，通过构建给定发电机的转子角与转子速度的相对偏差量的时间序列作为智能体输入，以及采用分布式架构以弥补本地信息收集不足的缺陷。紧急控制方面，文献[110]将 DDPG 算法应用于应对紧急情况的发电机跳闸控制中，考虑到单一时间断面的观测信息即发电机的功角、角速度、机械功率和电磁功率并不足以描述动态过程，将多个时刻的观测量一并作为智能体的输入信息；文献[111]在故障导致的

电压恢复问题中将各个时刻的节点电压幅值和剩余负荷量作为观测空间，由于观测无法反映系统动态特性，引入长短时记忆(long short-term memory, LSTM)网络处理具有强相关性的时序信息。

综合能源系统管理方面，文献[112]研究微电网中的分布式能源管理问题，考虑到各个智能体无法实现对全局信息的观测，采用分布式的多智能体控制方法缓解观测缺失的问题；文献[113]针对分布式电源优化调度面临的隐私保护和实时决策问题，提出了基于联邦强化学习的多智能体分布式协同优化策略，具有隐私保护的优势。信息安全方面，文献[114]采用 DQN 方法检测并防御电力系统中的数据完整性攻击，其中系统运营商不能直接观测到系统是否正常运行，只能根据观测变量即量测数据和状态估计结果的差值做出评估，为解决部分可观问题，引入观测变量的时间序列扩充作为增广后的状态；文献[115]将在线攻击/异常检测问题表述为 POMDP 问题，通过量测结果与对应的状态估计结果推断攻击是否进行，针对部分可观性以滑动时间窗内的观测值序列作为智能体的输入状态。

2.2.3 小结

电力系统的部分可观测性取决于多种因素，包括状态转移过程的自身特征、实际电力系统的某些特征量是否无法或很难量测等。一部分应用场景具有良好的完全可观测性，可以将其建模为 MDP，进一步通过各种强化学习方法解决；另一些控制优化场景则只能实现部分可观测，常见的解决方法包括将其建模成为 POMDP，或者引入时间序列增广的输入状态、采用具有记忆特性的网络结构或者深层特征提取结构等对部分可观测性进行针对性的处理，从而实现对深层时序数据的特征提取。

2.3 强化学习应用于电力系统优化与控制过程的最优性或次优性保证问题

强化学习应用于电力系统优化与控制过程的目的之一是在复杂应用场景中获得有一定性能保证的控制效果。尽管强化学习在电力系统各领域优化与控制中的一系列应用取得了不错的效果，但大多数应用场景使用的强化学习算法依然缺乏学习效果（例如收敛性以及最优性/次优性）的理论保证。一部分强化学习算法不具备较好的理论层面的收敛性保证，学习过程存在震荡甚至可能发散[116]；而如 TRPO/PPO 等具有持续优化理论保证的算法或具有约束的算法通常具有较好的学习性能，考虑到学习效果的改善，在有足够样本的前提下可达到至少次优的学习效果；一些表格型学习算法已经有收敛性和最优性的证明。

与此同时，强化学习的最优性或次优性要求与安全约束有密切联系。针对电力系统运行的安全约束，既可以通过构造奖励函数将安全约束以罚函数的形式作为软性约束，又可以通过修改优化准则等方式将约束条件采用拉格朗日对偶方法与目标函数结合。强化学习方法应用于实际电力系统场景的最优性/次优性的理论层面保证，依然是需要关注的重点问题之一。本节针对不同的算法对最优性或次优性保证进行分析与综述，表 3 列举了强化学习应用于电力系统优化与控制过程最优性或次优性保证的文献摘要[38,89,123-137]。

2.3.1 具备最优性或次优性保证的算法应用

根据智能体数目的不同，强化学习算法可分为单智能体强化学习(single agent reinforcement learning，SARL)和多智能体强化学习(multi agent reinforcement learning，MARL)这 2 大类。SARL 理论方面，以 TD 和 Q 学习为代表的一系列表格型强化学习方法具有较完备的收敛性和最优性证明[117–120]。同时，精确策略梯度算法在宽松的假设条件下可收敛到最优策略[57]。应用上述表格类或精确策略梯度算法的应用场景在理论上具有最优性或次优性保证。MARL 理论方面，Nash-Q learning[121]和 Friend-or-Foe Q-learning[122]为代表的一类多智能体算法在理论上可以收敛到纳什均衡。

SARL 应用方面，文献[123]比较了模型预测控制与拟合 Q 迭代算法(fitted Q-iteration，FQI)在时域有和无限情况下是如何解决最优控制问题的，给出了 2 者的次优性上限，并介绍了 2 种方法在电力系统振荡阻尼问题上的应用；文献[124]采用 Q 学习方法实现故障下的自适应发电机功率调节以缓解线路过载，在 IEEE 118 节点系统的验证表明了策略收敛到最优策略的邻域；文献[125]将电网无功电压优化控制建模成 MDP，利用强化学习中 Q 学习算法的渐进学习寻优能力优化地区电网无功电压控制策略，能够实时给出当前学习阶段下的最佳控制策略。

MARL 应用方面，文献[126]将多智能体 Q 学习(multi agent Q-learning)应用于多微电网的二次优化问题，分布的智能体基于全局奖励和本地奖励更新自身的 Q 函数表格，最终收敛到最优控制策略；文献[127]以多智能体架构优化暖通空调(heating，

表 3　强化学习应用于电力系统优化与控制过程最优性或次优性保证的文献摘要，
Table 3　Literature summary on optimality or sub-optimality guarantees of reinforcement learning methods applied to power system optimization and control scenarios

文献	学习算法	研究领域	研究场景	是否有最优性/次优性保证
文献[123]	FQI	频率控制	电力系统振荡阻尼	是
文献[124]	Q-learning	紧急控制	电力系统的 N-1 和 N-2 运行	是
文献[125]	Q-learning	电压控制	地区电网无功电压优化控制	是
文献[126]	Multi Agent Q-learning	优化调度	多微电网分布式二次优化控制	是
文献[127]	Multi Agent Q-learning	综合能源系统管理	含暖通空调系统的建筑能源管理	是
文献[128]	Multi Agent Q-learning	电力市场	综合能源市场交易优化	是
文献[129]	Constrained DDPG	优化调度	实时最优潮流求解	是
文献[130]	Constrained DDPG	电压控制	考虑稳定约束的配电网实时电压控制	是
文献[89]	Soft Actor-Critic	预防控制	电力系统静态安全预防控制	是
文献[131]	PPO	优化调度	考虑新能源不确定性的无功优化	是
文献[132]	Robust-ADP	频率控制	多机电力系统的分布式发电控制	是
文献[133]	Robust-ADP	综合能源系统管理	双馈感应发电机的无功功率控制	是
文献[134]	DQN/DDPG	电压控制	电力系统的自动电压控制	否
文献[38]	DDPG	电压控制	紧急低压减载	否
文献[135]	Actor Critic	电压控制	多端背靠背柔性直流系统电压控制	否
文献[136]	Multi Agent DDPG	优化调度	有功-无功协调控制	否
文献[137]	Multi Agent DDPG	电力市场	发电公司的投标策略分析	否

ventilation and air conditioning，HVAC)系统的控制与电力系统的规划，各控制器依据自身信息和全局奖励更新自身的 Q 表格和策略，仿真结果验证了算法的有效性；文献[128]基于多智能体 Nash Q 学习构建电-气综合能源市场多参与主体竞价博弈应用框架，基于 TD 算法更新各智能体的 Q 函数表格，所得策略具有次优性。

此外，通过向 DRL 的学习框架中添加约束项或限制策略搜索空间也可以增强强化学习的收敛性与最优性保证。如 TRPO 算法引入对目标函数的约束，利用拉格朗日对偶性改写目标函数，结合信赖域方法实现了学习效果的单调不减。文献[129]针对实时最优潮流问题提出了一种基于拉格朗日方法的深度强化学习方法，在目标函数中增加了针对运行条件的约束项，仿真结果验证了添加约束项的算法的收敛性和最优性；文献[130]在配电网实时电压控制问题中通过李雅普诺夫约束构造了一个有稳定性保证的智能体策略搜索子集，从而保证强化学习更新的稳定性和单调性；文献[89]基于 SAC 算法研究电力系统静态安全预防控制，在收敛速度与稳定性方面相较于 DDPG 算法更具优势；文献[131]研究基于具有性能保证的 PPO 算法的电网无功优化，策略更新具有一定的性能保证，仿真验证了所提方法相对于传统确定性优化算法更好的决策效果；文献[132]研究了鲁棒自适应动态规划(robust adaptive dynamic programming，Robust-ADP)方法在多机电力系统的在线学习控制的应用，提出的 Robust-ADP 框架可以得到具有次优性和鲁棒性的全局渐近稳定控制策略；文献[133]基于 ADP 的在线补充学习控制方法设计了一种优化和自适应双馈无功补充控制设计方法，以提高电力系统的暂态稳定性，其中采用基于最小二乘的策略迭代算法训练得到具有收敛性和稳定性保证的辅助控制器。

2.3.2　不具备最优性或次优性保证的算法应用

大部分具有函数逼近的强化学习算法并不具备严格的理论层面的收敛性或最优性保证，如常见的 DQN 或者 DDPG 算法均有发散的风险。因此，绝大部分强化学习算法应用于电力系统优化与控制的场景虽然能获得较好的效果，但依然缺乏理论层面的最优性或次优性的保证。

文献[134]提出了 1 种基于深度强化学习的电力系统自主控制框架，并以自动电压控制为例进行了 DQN 和 DDPG 算法的应用，2 种算法均表现出了良好的效果，但均没有收敛到理论最优解；文献[38]将 DDPG 算法应用于紧急低压减载场景中，仿真验证中智能体表现出了较好的控制效果，但 Critic 网络并未收敛到无偏估计；文献[135]采用 Actor-Critic 算法研究多端背靠背柔性直流系统电压控制，虽然算法表现出了良好的效果，但未能完全消除智能体策略收敛到局部最优或发散的风险，

只能在一定程度上降低发散风险；文献[136]基于多智能体深度确定性策略梯度(multi agent DDPG，MADDPG)建立电网分层有功-无功协调调度框架，仿真结果表明了策略的收敛性，但并未说明策略收敛到最优或者次优；文献[137]在对发电公司投标策略的仿真中验证了不同的参数设置会导致不同的收敛结果，不合理的超参数设置甚至会使所得策略发散。

近年来，强化学习基础理论界对 DRL 的适用性有一定研究，文献[138]用实验方法验证了在 DQN 中很少出现不收敛的情况，文献[139]则研究了目标网络在影响收敛性方面的作用，为目标网络稳定训练的传统观点提供了理论支持。因此，上述案例虽缺乏严格的理论保证，但仍获得了较好的应用效果。不过，将强化学习方法落地于实际电力系统情境的过程中，如何构建最优性/次优性的理论层面保证，依然是需要关注的重点问题之一。

2.3.3 小结

根据性能保证对强化学习方法的分类如图 3 所示。在状态和动作空间离散的场景下应用表格型强化学习算法或在代价敏感的场景中针对强化学习算法引入合适的约束条件并与目标函数相结合，可以提升强化学习算法的收敛性与最优性的理论保证。大部分强化学习算法并不具备严格的理论层面的收敛性和最优性/次优性保证，虽然电力系统各领域基于上述方法的若干应用取得了不错的效果，但是在应用落地过程中仍应对其安全性进行相应的约束，以及对所得模型的优化与控制效果进行更全面的评估。

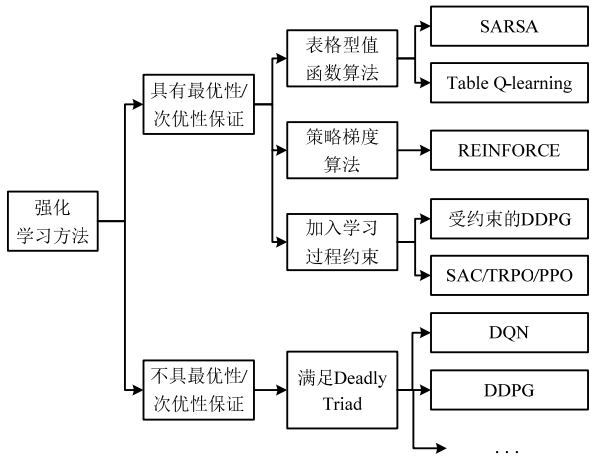


图 3 根据性能保证对强化学习方法的分类

Fig. 3 Classification of reinforcement learning methods according to optimality or sub-optimality guarantees

2.4 强化学习应用于电力系统优化与控制问题中的其他挑战

2.4.1 分布式强化学习的理论性能与扩展性

随着高比例可再生能源馈入电力系统导致电力系统中的分布式发电(distributed generation，DG)数目剧增，电力系统优化与控制场景的状态空间和动作空间维数急剧增加且动态变化。具有良好扩展性的分布式强化学习成为强化学习在电力系统的重要应用形式之一[140-144]。但由于缺乏针对分布式强化学习的理论性能保证(如分布式控制导致的环境非平稳)与维度指数增加带来的扩展性问题等[145-146]，分布式强化学习架构应用于电力系统的研究依然主要停留于应用层次的研究。文献[147]提出了一种具有理论性能保证的零阶分布式策略优化算法，并在多区域 HVAC 系统中验证了算法的可行性。文献[148]提出了一种利用网络结构的特性的可扩展强化学习框架，并证明了可扩展的演员-评论家方法可以获得次优策略。但目前为止仍缺乏对分布式强化学习应用于电力系统优化控制的理论基础研究。

2.4.2 离线仿真环境和在线实际系统的偏差

受电力系统建模与仿真的精度制约，离线仿真的环境往往与现实系统存在一定偏差。这一偏差将降低训练数据集的数据质量，进而影响训练得到的智能体的性能，甚至可能导致现实环境中的优化控制结果出现根本性偏差，对系统的安稳运行带来严重威胁[149]。在电力系统领域，文献[150]研究离线强化学习和在线强化学习应用于储能系统优化运行，并指出仿真环境与现实环境的差异会影响学习效果；文献[151]采用领域自适应方法学习同一电网不同运行水平的调度，采用策略迁移学习不同电网模型的调度。如何衡量仿真模型与实际系统的差异，构建适用于电力系统的考虑样本偏差的强化学习算法，是强化学习算法应用于实际电力系统需要解决的问题之一。

2.4.3 强化学习方法应用于电力系统的可解释性

作为一种数据驱动的方法，强化学习面临着可解释性方面的挑战[152-153]。考虑到电力系统优化与控制场景的代价敏感程度极高，对强化学习方法应用于电力系统的全过程进行解释、了解其在实际中的运作机理是强化学习方法落地于电力系统优化控制场景的关键挑战之一[154-155]。文献[156]根据 Deep-SHAP 方法对基于深度强化学习的电力系统

紧急控制应用提供合理的可解释模型，而文献[157]基于加权斜决策树算法对强化学习智能体训练得到的低压减载策略进行策略提取，说明其可解释性。但是，如何对复杂的决策模型提供合理可信的解释，同时不过多丢失关键信息，目前依然是阻碍强化学习落地的因素之一。

2.4.4　强化学习自身的理论局限性

强化学习在诸多复杂的电力系统优化控制场景中取得巨大成功的同时，也面临其自身存在的一些理论局限性，如学习效率偏低等。文献[158]说明了强化学习的学习效率较低的原因是网络的增量更新与较弱的归纳偏置。在电力系统控制优化方面，以文献[105,159]为例，上述应用场景均需要数万个运行场景对强化学习智能体进行训练才能达到较好的效果。近年来，强化学习理论界针对样本效率的问题也进行了相应的研究，文献[160]从提高环境的采样效率和提高已有样本的利用率 2 个方面对现有的研究成果进行了总结。文献[161]则从探索、优化、环境建模、经验迁移和抽象等方面讨论了减轻强化学习样本成本的可能方法。近年来理论界对强化学习的样本效率改进也提出了一系列的方法[162–165]。

3　强化学习应用于电力系统优化与控制领域的研究展望

虽然强化学习应用于电力系统依然面临前述的若干关键性问题与挑战，但强化学习方法依然具有广阔的应用前景。基于前述的若干关键挑战，本节提出了强化学习应用于电力系统优化与控制领域的几点研究展望。

3.1　强化学习与电力系统机理特性的融合

文献[166]提出将物理知识嵌入机器学习的若干实现途径，嵌入物理知识后的机器学习模型在增强机理性与可解释性的同时，获得了更好的应用效果。类似地，将电力系统自身的物理特性等信息作为先验知识与强化学习方法有机融合，一方面可以增强强化学习方法应用过程的可解释性，另一方面也能够基于电力系统自身的某些有别于其他应用场景的特性(例如数据有限和具有一部分环境的先验信息等)对应用场景进行针对性的马尔可夫建模，构建更适用于电力系统优化与控制的强化学习方法。

3.2　应用于部分可观环境下的强化学习

新一代电力系统的感知能力极大提升，但仍存在大量部分可观测的优化与控制应用场景，针对部分可观场景下强化学习适用性与学习效果的研究依然有重要的意义。针对部分可观测的以微分方程描述的动态系统的最优控制，强化学习理论界已有相关研究[167-168]。具有记忆单元的深度网络与 DRL 的结合在部分可观场景中的应用也取得了良好的效果[96,169-171]。将上述研究引入电力系统优化与控制过程，有利于提升智能体对数据特征的深度挖掘能力，推动强化学习在部分可观场景中的应用。

3.3　针对代价敏感场景的安全强化学习

电力系统是代价高度敏感的系统，如果决策产生失误，有可能导致严重的后果。虽然目前几乎所有的应用于电力系统的强化学习方法都基于离线的仿真模型进行学习，但前述的环境差异同样会引入风险因素。针对上述问题，强化学习界提出了“安全强化学习”的概念[172-174]，即在满足一定约束的前提下使期望收益最大化。针对基于离线样本的离线强化学习方法存在的分布偏差问题，近年来也已经有一些研究[175-177]。此外，鲁棒自适应动态规划(robust approximate/adaptive dynamic programming, Robust-ADP)方法在电力系统优化控制中也已经有一些应用[132-133]。在强化学习应用于电力系统的过程中引入上述概念方法，将有助于提升优化与控制的鲁棒性，同时使得优化控制策略具有一定的安全性和最优性保证。

3.4　强化学习与其他人工智能方法的有机结合

随着电力系统向更主动、更灵活、更智能的智能电网过渡，多种人工智能方法在电力系统优化控制中展现出良好的效果[178-179]。目前已经有一些将强化学习方法与其他人工智能方法结合的应用，如利用自编码器进行特征挖掘并用于优化强化学习控制器参数[72]，或利用具有一定可解释性的人工智能模型如加权斜决策树对强化学习模型进行控制策略提取[157]，或在强化学习算法中引入对抗网络实现长期智能微电网的发电控制[180]。将强化学习方法与其他人工智能方法进行有机结合，可以相辅相成，有助于增强智能体对复杂特征的挖掘能力以及对未知场景的适应能力，同时可以增强其可解释性。

4　结论

新一代电力系统的“双高”特性对电力系统一些领域的优化与控制带来了严峻的挑战。有赖于电力系统广域信息感知技术的发展，基于数据驱动的

强化学习方法在电力系统各领域的优化与控制场景中的应用表现出良好的效果，但在强化学习方法落地于实际电力系统应用的过程中依然存在一系列关键问题与挑战。本文简述了强化学习的基础理论与研究现状，并对强化学习应用于电力系统优化与控制的 3 大关键问题进行综述，即马尔科夫属性是否满足、应用场景是否完全可观、学习算法是否理论最优或次优，并提出了针对强化学习落地于电力系统实际应用的研究展望。明确强化学习方法适用于怎样的场景，以及应用强化学习方法的预期效果如何，是强化学习应用于电力系统各领域优化与控制过程亟待解决的问题。

参考文献

[1] 周孝信，陈树勇，鲁宗相，等．能源转型中我国新一代电力系统的技术特征[J]．中国电机工程学报，2018，38(7)：1893-1904．
ZHOU Xiaoxin，CHEN Shuyong，LU Zongxiang，et al．Technology features of the new generation power system in China[J]．Proceedings of the CSEE，2018，38(7)：1893-1904(in Chinese)．

[2] 鲁宗相，李海波，乔颖．含高比例可再生能源电力系统灵活性规划及挑战[J]．电力系统自动化，2016，40(13)：147-158．
LU Zongxiang，LI Haibo，QIAO Ying．Power system flexibility planning and challenges considering high proportion of renewable energy[J]．Automation of Electric Power Systems，2016，40(13)：147-158(in Chinese)．

[3] 马宁宁，谢小荣，贺静波，等．高比例新能源和电力电子设备电力系统的宽频振荡研究综述[J]．中国电机工程学报，2020，40(15)：4720-4731．
MA Ningning，XIE Xiaorong，HE Jingbo，et al．Review of wide-band oscillation in renewable and power electronics highly integrated power systems[J]．Proceedings of the CSEE，2020，40(15)：4720-4731(in Chinese)

[4] 张宁，马国明，关永刚，等．全景信息感知及智慧电网[J]．中国电机工程学报，2021，41(4)：1274-1283．
ZHANG Ning，MA Guoming，GUAN Yonggang，et al．Panoramic information perception and intelligent grid[J]．Proceedings of the CSEE，2021，41(4)：1274-1283(in Chinese)．

[5] 郭剑波．新型电力系统面临的挑战以及有关机制思考[J]．中国电力企业管理，2021(25)：8-11．
GUO Jianbo．Challenges of the new power system and related mechanism considerations[J]．China Power Enterprise Management，2021(25)：8-11(in Chinese)．

[6] KAELBLING L P，LITTMAN M L，MOORE A W．Reinforcement learning：a survey[J]．Journal of Artificial Intelligence Research，1996，4(1)：237-285．

[7] SUTTON R S，BARTO A G．Reinforcement learning：An introduction[M]．Cambridge：MIT Press，2018．

[8] 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．

[9] MOUSAVI S S，SCHUKAT M，HOWLEY E．Deep reinforcement learning：An overview[C]//．Proceedings of SAI Intelligent Systems Conference(IntelliSys) 2016．Cham：Springer，2018：426-440．

[10] LI Yuxi．Deep reinforcement learning：An overview[J]．arXiv：1701.07274，2017．

[11] LAMPLE G，CHAPLOT D S．Playing FPS games with deep reinforcement learning[C]//Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence．San Francisco：AAAI Press，2017．

[12] SILVER D，HUBERT T，SCHRITTWIESER J，et al．A general reinforcement learning algorithm that masters chess，shogi，and Go through self-play[J]．Science，2018，362(6419)：1140-1144．

[13] ZHA Daochen，XIE Jingru，MA Wenye，et al．DouZero：mastering DouDizhu with self-play deep reinforcement learning[C]//Proceedings of the 38th International Conference on Machine Learning．Online：PMLR，2021：12333-12344．

[14] TUNYASUVUNAKOOL K，ADLER J，WU Z，et al．Highly accurate protein structure prediction for the human proteome[J]．Nature，2021，596(7873)：590-596．

[15] DEGRAVE J，FELICI F，BUCHLI J，et al．Magnetic control of tokamak plasmas through deep reinforcement learning[J]．Nature，2022，602(7897)：414-419．

[16] 余涛，周斌，甄卫国．强化学习理论在电力系统中的应用及展望[J]．电力系统保护与控制，2009，37(14)：122-128．
YU Tao，ZHOU Bin，ZHEN Weiguo．Application and development of reinforcement learning theory in power systems[J]．Power System Protection and Control，2009，37(14)：122-128(in Chinese)．

[17] GLAVIC M，FONTENEAU R，ERNST D．Reinforcement learning for electric power system decision and control：past considerations and perspectives[J]．IFAC-PapersOnLine，2017，50(1)：6918-6927．

[18] ZHANG Dongxia，HAN Xiaoqing，DENG Chunyu．Review on the research and practice of deep learning and reinforcement learning in smart grids[J]．CSEE Journal of Power and Energy Systems，2018，4(3)：362-370．

[19] GLAVIC M．(Deep) Reinforcement learning for electric power system control and related problems：A short

review and perspectives[J]. Annual Reviews in Control，2019，48：22-35.

[20] 孙毅，刘迪，李彬，等. 深度强化学习在需求响应中的应用[J]. 电力系统自动化，2019，43(5)：183-191.
SUN Yi，LIU Di，LI Bin，et al. Application of deep reinforcement learning in demand response[J] . Automation of Electric Power Systems，2019，43(5)：183-191(in Chinese).

[21] 范士雄，李立新，王松岩，等. 人工智能技术在电网调控中的应用研究[J]. 电网技术，2020，44(2)：401-411.
FAN Shixiong，LI Lixin，WANG Songyan，et al. Application analysis and exploration of artificial intelligence technology in power grid dispatch and control[J]. Power System Technology，2020，44(2)：401-411(in Chinese).

[22] ZHANG Zidong，ZHANG Dongxia，QIU R C. Deep reinforcement learning for power system applications：An overview[J] . CSEE Journal of Power and Energy Systems，2020，6(1)：213-225.

[23] 宋鹏飞，杨宁，崔承刚，等. 深度强化学习应用于电力系统控制研究综述[J]. 现代计算机，2021(1)：39-44.
SONG Pengfei，YANG Ning，CUI Chenggang，et al. Survey of the application of deep reinforcement learning in power system control[J]. Modern Computer，2021(1)：39-44(in Chinese).

[24] 熊珞琳，毛帅，唐漾，等. 基于强化学习的综合能源系统管理综述[J]. 自动化学报，2021，47(10)：2321-2340.
XIONG Luolin，MAO Shuai，TANG Yang，et al. Reinforcement learning based integrated energy system management：A survey[J]. Acta Automatica Sinica，2021，47(10)：2321-2340(in Chinese)

[25] CHEN Xin，QU Guannan，TANG Yujie，et al. Reinforcement learning for selective key applications in power systems：Recent advances and future challenges[J]. IEEE Transactions on Smart Grid，2022，13(4)：2935-2958.

[26] 张有兵，林一航，黄冠弘，等. 深度强化学习在微电网系统调控中的应用综述[J]. 电网技术，2023，47(7)：2774-2788.
ZHANG Youbing，LIN Yihang，HUANG Guanhong，et al. Review on applications of deep reinforcement learning in regulation of microgrid systems[J]. Power System Technology，2023，47(7)：2774-2788(in Chinese).

[27] LEWIS F L，LIU Derong. Reinforcement learning and approximate dynamic programming for feedback control[M]. Hoboken：Wiley-IEEE Press，2013.

[28] LU Chao，SI J，WU Xiaochen，et al. Approximate dynamic programming coordinated control in multi-infeed HVDC power system[C]//2006 IEEE PES Power Systems Conference and Exposition. Atlanta：IEEE，2006：2131-2135.

[29] LIANG Jiaqi，VENAYAGAMOORTHY G K，HARLEY R G. Wide-area measurement based dynamic stochastic optimal power flow control for smart grids with high variability and uncertainty[J]. IEEE Transactions on Smart Grid，2012，3(1)：59-69.

[30] YU Miao，LU Chao，LIU Yongjun. Direct heuristic dynamic programming method for power system stability enhancement[C]//2014 American Control Conference. Portland：IEEE，2014：747-752.

[31] TANG Yufei，HE Haibo，WEN Jinyu，et al. Power system stability control for a wind farm based on adaptive dynamic programming[J]. IEEE Transactions on Smart Grid，2015，6(1)：166-177.

[32] XI Lei，CHEN Jianfeng，HUANG Yuehua，et al. Smart generation control based on multi-agent reinforcement learning with the idea of the time tunnel[J]. Energy，2018，153：977-987.

[33] YAN Ziming，XU Yan. Data-driven load frequency control for stochastic power systems：A deep reinforcement learning method with continuous action search[J]. IEEE Transactions on Power Systems，2019，34(2)：1653-1656.

[34] ZHANG Jingyi，LUO Yonglong，WANG Boya，et al. Deep reinforcement learning for load shedding against short-term voltage instability in large power systems[J]. IEEE Transactions on Neural Networks and Learning Systems，2021：1-12. doi: 10.1109/TNNLS.2021.3121757.

[35] 邓清唐，胡丹尔，蔡田田，等. 基于多智能体深度强化学习的配电网无功优化策略[J]. 电工电能新技术，2022，41(2)：10-20.
DENG Qingtang，HU Dan’er，CAI Tiantian，et al. Reactive power optimization strategy of distribution network based on multi agent deep reinforcement learning[J] . Advanced Technology of Electrical Engineering and Energy，2022，41(2)：10-20(in Chinese).

[36] SUN Jian，QI Guanqiu，MAZUR N，et al. Structural scheduling of transient control under energy storage systems by sparse-promoting reinforcement learning[J]. IEEE Transactions on Industrial Informatics，2022，18(2)：744-756.

[37] ZHANG Jingyi，LU Chao，FANG Chen，et al. Load shedding scheme with deep reinforcement learning to improve short-term voltage stability[C]//2018 IEEE Innovative Smart Grid Technologies - Asia(ISGT Asia). Singapore：IEEE，2018：13-18.

[38] LI Jian，CHEN Sheng，WANG Xinying，et al. Load shedding control strategy in power grid emergency state

based on deep reinforcement learning[J]. CSEE Journal of Power and Energy Systems，2022，8(4)：1175-1182.

[39] 刘威，张东霞，王新迎，等. 基于深度强化学习的电网紧急控制策略研究[J]. 中国电机工程学报，2018，38(1)：109-119.
LIU Wei，ZHANG Dongxia，WANG Xinying，et al. A decision making strategy for generating unit tripping under emergency circumstances based on deep reinforcement learning[J]. Proceedings of the CSEE，2018，38(1)：109-119(in Chinese).

[40] VU T L，MUKHERJEE S，YIN T，et al. Safe reinforcement learning for emergency load shedding of power systems[C]//2021 IEEE Power & Energy Society General Meeting(PESGM). Washington：IEEE，2021：1-5.

[41] DULAC-ARNOLD G，LEVINE N，MANKOWITZ D J，et al. Challenges of real-world reinforcement learning：Definitions，benchmarks and analysis[J]. Machine Learning，2021，110(9)：2419-2468.

[42] SUTTON R S. The quest for a common model of the intelligent decision maker[J]. arXiv：2202.13252，2022.

[43] 钱敏平，龚光鲁. 随机过程论[M]. 2 版. 北京：北京大学出版社，1997.
QIAN Minping，GONG Guanglu. The theory of stochastic processes[M]. 2nd Edition. Beijing：Peking University Press，1997(in Chinese).

[44] ARULKUMARAN K，DEISENROTH M P，BRUNDAGE M，et al. Deep reinforcement learning：A brief survey[J]. IEEE Signal Processing Magazine，2017，34(6)：26-38.

[45] RUMMERY G A，NIRANJAN M. On-line Q-learning using connectionist systems[M]. Cambridge：University of Cambridge，1994.

[46] WATKINS C J C H，DAYAN P. Q-learning[J]. Machine Learning，1992，8(3)：279-292.

[47] MNIH V，KAVUKCUOGLU K，SILVER D，et al. Playing atari with deep reinforcement learning[J]. arXiv：1312.5602，2013.

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

[49] SUTTON R S，MCALLESTER D，SINGH S，et al. Policy gradient methods for reinforcement learning with function approximation[C]//Proceedings of the 12th International Conference on Neural Information Processing Systems. Denver：MIT Press，1999：1057-1063.

[50] WILLIAMS R J. Simple statistical gradient-following algorithms for connectionist reinforcement learning[J]. Machine Learning，1992，8(3)：229-256.

[51] KONDA V R，TSITSIKLIS J N. Actor-Critic Algorithms[C]//Proceedings of the 13th International Conference on Neural Information Processing Systems. Denver：MIT Press，2000.

[52] LILLICRAP T P，HUNT J J，PRITZEL A，et al. Continuous control with deep reinforcement learning[J]. Computer Science，2015，8(6)：A187.

[53] TAYLOR C R. Applications of dynamic programming to agricultural decision problems[M]. Boca Raton：CRC Press，2019：1-10.

[54] XU Xin，ZUO Lei，HUANG Zhenhua. Reinforcement learning algorithms with function approximation：Recent advances and applications[J]. Information Sciences，2014，261：1-31.

[55] PARR R，LI Lihong，TAYLOR G，et al. An analysis of linear models，linear value-function approximation，and feature selection for reinforcement learning[C]//Proceedings of the 25th International Conference on Machine Learning. Helsinki：Association for Computing Machinery，2008：752-759.

[56] MELO F S，MEYN S P，RIBEIRO M I. An analysis of reinforcement learning with function approximation[C]//Proceedings of the 25th International Conference on Machine Learning. Helsinki：Association for Computing Machinery，2008：664-671.

[57] AGARWAL A，KAKADE S M，LEE J D，et al. On the theory of policy gradient methods：Optimality，approximation，and distribution shift[J]. Journal of Machine Learning Research，2021，22(1)：98.

[58] NIAN Rui，LIU Jinfeng，HUANG Biao. A review on reinforcement learning：Introduction and applications in industrial process control[J]. Computers & Chemical Engineering，2020，139：106886.

[59] BAIRD L. Residual algorithms：Reinforcement learning with function approximation[M]//PRIEDITIS A，RUSSELL S. Machine Learning Proceedings 1995. San Francisco：Morgan Kaufmann，1995：30-37.

[60] HENDERSON P，ISLAM R，BACHMAN P，et al. Deep reinforcement learning that matters[J]. Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence. New Orleans：AAAI Press，2018：392.

[61] TSITSIKLIS J N，VAN ROY B. An analysis of temporal-difference learning with function approximation[J]. IEEE Transactions on Automatic Control，1997，42(5)：674-690.

[62] XU Pan，GU Quanquan. A finite-time analysis of q-learning with neural network function approximation

[C]//Proceedings of the 37th International Conference on Machine Learning．Online：PMLR，2020：978．

[63] AGARWAL A，KAKADE S M，LEE J D，et al．Optimality and approximation with policy gradient methods in markov decision processes[C]//Proceedings of Thirty Third Conference on Learning Theory．Graz：PMLR，2020：64-66．

[64] FAN Jianqing，WANG Zhaoran，XIE Yuchen，et al．A theoretical analysis of deep Q-learning[C]//Proceedings of the 2nd Conference on Learning for Dynamics and Control．Berkeley：PMLR，2020：486-489．

[65] WANG Ruosong，SALAKHUTDINOV R R，YANG L．Reinforcement learning with general value function approximation：Provably efficient approach via bounded eluder dimension[C]//Proceedings of the 34th International Conference on Neural Information Processing Systems．Online：Curran Associates Inc.，2020：6123－6135．

[66] YANG Zhuoran，JIN Chi，WANG Zhaoran，et al．On function approximation in reinforcement learning：optimism in the face of large state spaces[J]．arXiv：2011.04622，2020．

[67] AMANI S，THRAMPOULIDIS C，YANG Lin．Safe reinforcement learning with linear function approximation [C]//Proceedings of the 38th International Conference on Machine Learning．Online：PMLR，2021：243-253．

[68] LEE J，PACCHIANO A，MUTHUKUMAR V，et al．Online model selection for reinforcement learning with function approximation[C]//Proceedings of the 24th International Conference on Artificial Intelligence and Statistics．Online：PMLR，2021：3340-3348．

[69] 习伟，李鹏，李鹏，等．基于深度强化学习的分布式电源就地自适应电压控制方法[J]．电力系统自动化，2022，46(22)：25-31．
XI Wei，LI Peng，LI Peng，et al．Adaptive local voltage control method for distributed generator based on deep reinforcement learning[J]．Automation of Electric Power Systems，2022，46(22)：25-31(in Chinese)．

[70] YIN Linfei，ZHANG Chenwei，WANG Yaoxiong，et al．Emotional deep learning programming controller for automatic voltage control of power systems[J]．IEEE Access，2021，9：31880-31891．

[71] HOSSAIN R R，HUANG Qiuhua，HUANG Renke．Graph convolutional network-based topology embedded deep reinforcement learning for voltage stability control[J]．IEEE Transactions on Power Systems，2021，36(5)：4848-4851．

[72] 杨丽，孙元章，徐箭，等．基于在线强化学习的风电系统自适应负荷频率控制[J]．电力系统自动化，2020，44(12)：74-83．
YANG Li，SUN Yuanzhang，XU Jian，et al．Adaptive load frequency control of wind power system based on online reinforcement learning[J]．Automation of Electric Power Systems，2020，44(12)：74-83(in Chinese)．

[73] YAN Ziming，XU Yan．A multi-agent deep reinforcement learning method for cooperative load frequency control of a multi-area power system[J]．IEEE Transactions on Power Systems，2020，35(6)：4599-4608．

[74]CHEN Chunyu，CUI Mingjian，LI Fangxing，et al．Model-free emergency frequency control based on reinforcement learning[J]．IEEE Transactions on Industrial Informatics，2021，17(4)：2336-2346．

[75] 李琦，乔颖，张宇精．配电网持续无功优化的深度强化学习方法[J]．电网技术，2020，44(4)：1473-1480．
LI Qi，QIAO Ying，ZHANG Yujing．Continuous reactive power optimization of distribution network using deep reinforcement learning[J]．Power System Technology，2020，44(4)：1473-1480(in Chinese)．

[76] ZHANG Ying，WANG Xinan，WANG Jianhui，et al．Deep reinforcement learning based volt-var optimization in smart distribution systems[J]．IEEE Transactions on Smart Grid，2021，12(1)：361-371．

[77] 邱高，刘友波，许立雄，等．基于深度确定性策略梯度的电网断面极限传输能力动态趋优控制[J]．中国电机工程学报，2021，41(15)：5128-5138．
QIU Gao，LIU Youbo，XU Lixiong，et al．A deep deterministic policy gradient based-dynamic optimizing control for power system total transfer capability[J]．Proceedings of the CSEE，2021，41(15)：5128-5138(in Chinese)．

[78] 彭刘阳，孙元章，徐箭，等．基于深度强化学习的自适应不确定性经济调度[J]．电力系统自动化，2020，44(9)：33-42．
PENG Liuyang，SUN Yuanzhang，XU Jian，et al．Self-adaptive uncertainty economic dispatch based on deep reinforcement learning[J]．Automation of Electric Power Systems，2020，44(9)：33-42(in Chinese)．

[79] 杨挺，赵黎媛，刘亚闯，等．基于深度强化学习的综合能源系统动态经济调度[J]．电力系统自动化，2021，45(5)：39-47．
YANG Ting，ZHAO Liyuan，LIU Yachuang，et al．Dynamic economic dispatch for integrated energy system based on deep reinforcement learning[J]．Automation of Electric Power Systems，2021，45(5)：39-47(in Chinese)．

[80] DUAN Jiajun，YI Zhehan，SHI Di，et al．Reinforcement-learning-based optimal control of hybrid energy storage systems in hybrid AC－DC microgrids[J]．IEEE Transactions on Industrial Informatics，2019，15(9)：

5355-5364.

[81] 陈亭轩，徐潇源，严正，等. 基于深度强化学习的光储充电站储能系统优化运行[J]. 电力自动化设备，2021，41(10)：90-98.
CHEN Tingxuan，XU Xiaoyuan，YAN Zheng，et al. Optimal operation based on deep reinforcement learning for energy storage system in photovoltaic-storage charging station[J]. Electric Power Automation Equipment，2021，41(10)：90-98(in Chinese).

[82] LU Renzhi，HONG S H，ZHANG Xiongfeng. A dynamic pricing demand response algorithm for smart grid：Reinforcement learning approach[J]. Applied Energy，2018，220：220-230.

[83] DU Yan，LI Fangxing. Intelligent multi-microgrid energy management based on deep neural network and model-free reinforcement learning[J]. IEEE Transactions on Smart Grid，2020，11(2)：1066-1076.

[84] 邓祥力，王伟，刘世明. 基于博弈强化学习的电网故障序列搜索及防御策略研究[J]. 电网技术，2021，45(12)：4856-4867.
DENG Xiangli，WANG Wei，LIU Shiming. Research on searching of fault sequence and defense strategy in power grid based on game reinforcement learning[J]. Power System Technology，2021，45(12)：4856-4867(in Chinese).

[85] ROBERTS C，NGO S T，MILESI A，et al. Deep Reinforcement learning for DER cyber-attack mitigation [C]//2020 IEEE International Conference on Communications，Control，and Computing Technologies for Smart Grids(SmartGridComm). Tempe：IEEE，2020：1-7.

[86] 曾令康，姚伟，艾小猛，等. 基于双 Q 学习的考虑暂态稳定约束的电网薄弱线路辨识[J]. 中国电机工程学报，2020，40(8)：2429-2440.
ZENG Lingkang，YAO Wei，AI Xiaomeng，et al. Double Q-learning based identification of weak lines in power grid considering transient stability constraints[J]. Proceedings of the CSEE，2020，40(8)：2429-2440(in Chinese).

[87] 王甜婧，汤涌，郭强，等. 基于知识经验和深度强化学习的大电网潮流计算收敛自动调整方法[J]. 中国电机工程学报，2020，40(8)：2396-2405.
WANG Tianjing，TANG Yong，GUO Qiang，et al. Automatic adjustment method of power flow calculation convergence for large-scale power grid based on knowledge experience and deep reinforcement learning [J]. Proceedings of the CSEE，2020，40(8)：2396-2405(in Chinese).

[88] 李京，刘道伟，安军，等. 基于强化学习理论的静态电压稳定裕度评估[J]. 中国电机工程学报，2020，40(16)：5136-5147.
LI Jing，LIU Daowei，AN Jun，et al. Static voltage stability margin assessment based on reinforcement learning theory[J]. Proceedings of the CSEE，2020，40(16)：5136-5147(in Chinese).

[89] 李柏堉，赵津蔓，韩肖清，等. 基于双智能体深度强化学习的电力系统静态安全预防控制方法[J]. 中国电机工程学报，2023，43(5)：1818-1830.
LI Baiyu，ZHAO Jinman，HAN Xiaoqing，et al. Static security oriented preventive control of power system based on double deep reinforcement learning[J]. Proceedings of the CSEE，2023，43(5)：1818-1830(in Chinese).

[90] 王渝红，胡胜杰，宋雨妍，等. 基于强化学习理论的输电网扩展规划方法[J]. 电网技术，2021，45(7)：2829-2838.
WANG Yuhong，HU Shengjie，SONG Yuyan，et al. Transmission expansion planning based on reinforcement learning[J]. Power System Technology，2021，45(7)：2829-2838(in Chinese).

[91] 孟安波，许炫淙，陈嘉铭，等. 基于强化学习和组合式深度学习模型的超短期光伏功率预测[J]. 电网技术，2021，45(12)：4721-4728.
MENG Anbo，XU Xuancong，CHEN Jiaming，et al. Ultra short term photovoltaic power prediction based on reinforcement learning and combined deep learning model[J]. Power System Technology，2021，45(12)：4721-4728(in Chinese).

[92] WANG Xinan，WANG Yishen，SHI Di，et al. Two-stage WECC composite load modeling：A double deep Q-learning networks approach[J]. IEEE Transactions on Smart Grid，2020，11(5)：4331-4344.

[93] 仲悟之，汤涌. 电力系统微分代数方程模型的暂态电压稳定性分析[J]. 中国电机工程学报，2010，30(25)：10-16.
ZHONG Wuzhi，TANG Yong. Transient voltage stability analysis of differential-algebra equation in power system [J]. Proceedings of the CSEE，2010，30(25)：10-16(in Chinese).

[94] BEVRANI H，HIYAMA T. Intelligent automatic generation control[M]. Boca Raton：CRC Press，2017.

[95] MONAHAN G E. State of the art—a survey of partially observable markov decision processes：Theory，models，and algorithms[J]. Management Science，1982，28(1)：1-16.

[96] HAUSKNECHT M J，STONE P. Deep recurrent Q-learning for partially observable MDPs[C]//2015 AAAI Fall Symposium. Arlington：AAAI Press，2015.

[97] SHANI G，PINEAU J，KAPLOW R. A survey of

point-based POMDP solvers[J]. Autonomous Agents and Multi-Agent Systems，2013，27(1)：1-51.

[98] SINGH S，JAMES M R，RUDARY M R. Predictive state representations：A new theory for modeling dynamical systems[C]//Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. Banf：AUAI Press，2012.

[99] LITTMAN M L. A tutorial on partially observable Markov decision processes[J]. Journal of Mathematical Psychology，2009，53(3)：119-125.

[100] RAO R P N. Decision making under uncertainty：A neural model based on partially observable markov decision processes[J]. Frontiers in Computational Neuroscience，2010，4：146.

[101] KURNIAWATI H，YADAV V. An online POMDP solver for uncertainty planning in dynamic environment[M]// INABA M，CORKE P. Robotics Research. Cham：Springer，2016：611-629.

[102] HADIDI R，JEYASURYA B. Reinforcement learning based real-time wide-area stabilizing control agents to enhance power system stability[J]. IEEE Transactions on Smart Grid，2013，4(1)：489-497.

[103] YANG Qiuling，WANG Gang，SADEGHI A，et al. Two-timescale voltage control in distribution grids using deep reinforcement learning[J]. IEEE Transactions on Smart Grid，2020，11(3)：2313-2323.

[104] 王之伟，陆晓，刁瑞盛，等. 基于深度强化学习的电网自主控制与决策技术[J]. 电力工程技术，2020，39(6)：34-43.
WANG Zhiwei，LU Xiao，DIAO Ruisheng，et al. Deep-reinforcement-learning based autonomous control and decision making for power systems[J]. Electric Power Engineering Technology，2020，39(6)：34-43(in Chinese).

[105] ROCCHETTA R，BELLANI L，COMPARE M，et al. A reinforcement learning framework for optimal operation and maintenance of power grids[J]. Applied Energy，2019，241：291-301.

[106] 于一潇，杨佳峻，杨明，等. 基于深度强化学习的风电场储能系统预测决策一体化调度[J]. 电力系统自动化，2021，45(1)：132-140.
YU Yixiao，YANG Jiajun，YANG Ming，et al. Prediction and decision integrated scheduling of energy storage system in wind farm based on deep reinforcement learning[J]. Automation of Electric Power Systems，2021，45(1)：132-140(in Chinese).

[107] 张自东，邱才明，张东霞，等. 基于深度强化学习的微电网复合储能协调控制方法[J]. 电网技术，2019，43(6)：1914-1921.
ZHANG Zidong，QIU Caiming，ZHANG Dongxia，et al. A coordinated control method for hybrid energy storage system in microgrid based on deep reinforcement learning[J]. Power System Technology，2019，43(6)：1914-1921(in Chinese).

[108] CAO Junwei，ZHANG Wanlu，XIAO Zeqing，et al. Reactive power optimization for transient voltage stability in energy internet via deep reinforcement learning approach[J]. Energies，2019，12(8)：1556.

[109] 范培潇，柯松，杨军，等. 基于改进多智能体深度确定性策略梯度的多微网负荷频率协同控制策略[J]. 电网技术，2022，46(9)：3504-3514.
FAN Peixiao，KE Song，YANG Jun，et al. Load frequency coordinated control strategy of multi-microgrid based on improved MA-DDPG[J]. Power System Technology，2022，46(9)：3504-3514(in Chinese).

[110] LIN Bilin，WANG Huaiyuan，ZHANG Yang，et al. Real-time power system generator tripping control based on deep reinforcement learning[J]. International Journal of Electrical Power & Energy Systems，2022，141：108127.

[111] HUANG Renke，CHEN Yujiao，YIN Tianzhixi，et al. Learning and fast adaptation for grid emergency control via deep meta reinforcement learning[J]. IEEE Transactions on Power Systems，2022，37(6)：4168-4178.

[112] WEI Chun，ZHANG Zhe，QIAO Wei，et al. An adaptive network-based reinforcement learning method for MPPT control of PMSG wind energy conversion systems[J]. IEEE Transactions on Power Electronics，2016，31(11)：7837-7848.

[113] 蒲天骄，杜帅，李烨，等. 基于联邦强化学习的分布式电源协同优化策略[J/OL]. 电力系统自动化，2022：1-12[2022-12-08]. http://kns.cnki.net/kcms/detail/32.1180.tp.20221025.1937.013.html.
PU Tianjiao，DU Shuai，LI Ye，et al. Collaborative optimization dispatch strategy of distributed generator based on federated reinforcement learning[J/OL]. Automation of Electric Power Systems，2022：1-12[2022-12-08]. http://kns.cnki.net/kcms/detail/32.1180.tp.20221025.1937.013.html(in Chinese).

[114] AN Dou，YANG Qingyu，LIU Wenmao，et al. Defending against data integrity attacks in smart grid：a deep reinforcement learning-based approach[J]. IEEE Access，2019，7：110835-110845.

[115] KURT M N，OGUNDIJO O，LI Chong，et al. Online cyber-attack detection in smart grid：A reinforcement learning approach[J]. IEEE Transactions on Smart Grid，

2019，10(5)：5174-5185.

[116]FAIRBANK M，ALONSO E．The divergence of reinforcement learning algorithms with value-iteration and function approximation[C]//The 2012 International Joint Conference on Neural Networks(IJCNN)．Brisbane：IEEE，2012：1-8.

[117]TSITSIKLIS J N，VAN ROY B．Analysis of temporal-difference learning with function approximation[C]//Proceedings of the 9th International Conference on Neural Information Processing Systems．Denver：MIT Press，1996.

[118] SINGH S，JAAKKOLA T，LITTMAN M L，et al．Convergence results for single-step on-policy reinforcement-learning algorithms[J]．Machine Learning，2000，38(3)：287-308.

[119] MELO F S．Convergence of Q-learning：A simple proof[R]．Lisboa：Institute of Systems and Robotics，Instituto Superior Técnico，2001：1-4.

[120] SCHOKNECHT R．Optimality of reinforcement learning algorithms with linear function approximation[C]//Proceedings of the 15th International Conference on Neural Information Processing Systems．Cambridge：MIT Press，2002.

[121] HU Junling，WELLMAN M P．Multiagent reinforcement learning：theoretical framework and an algorithm[C]//Proceedings of the Fifteenth International Conference on Machine Learning．San Francisco，CA，USA：Morgan Kaufmann Publishers Inc．，1998：242-250.

[122] LITTMAN M L．Friend-or-foe Q-learning in general-sum games[C]//Proceedings of the Eighteenth International Conference on Machine Learning．San Francisco，CA，USA：Morgan Kaufmann Publishers Inc.，2001：322-328.

[123] ERNST D，GLAVIC M，CAPITANESCU F，et al．Reinforcement learning versus model predictive control：A comparison on a power system problem[J]．IEEE Transactions on Systems，Man，and Cybernetics，Part B(Cybernetics)，2009，39(2)：517-529.

[124] ZARRABIAN S，BELKACEMI R，BABALOLA A A．Reinforcement learning approach for congestion management and cascading failure prevention with experimental application[J]．Electric Power Systems Research，2016，141：179-190.

[125] 刁浩然，杨明，陈芳，等．基于强化学习理论的地区电网无功电压优化控制方法[J]．电工技术学报，2015，30(12)：408-414.
DIAO Haoran，YANG Ming，CHEN Fang，et al．Reactive power and voltage optimization control approach of the regional power grid based on reinforcement learning theory[J]．Transactions of China Electrotechnical Society，2015，30(12)：408-414(in Chinese).

[126] 沈珺，柳伟，李虎成，等．基于强化学习的多微电网分布式二次优化控制[J]．电力系统自动化，2020，44(5)：198-206.
SHEN Jun，LIU Wei，LI Hucheng，et al．Reinforcement learning based distributed secondary optimal control for multiple microgrids[J]．Automation of Electric Power Systems，2020，44(5)：198-206(in Chinese).

[127] HAO Jun，GAO D W，ZHANG J J．Reinforcement learning for building energy optimization through controlling of central HVAC system[J]．IEEE Open Access Journal of Power and Energy，2020，7：320-328.

[128] 孙庆凯，王小君，王怡，等．基于多智能体 Nash-Q 强化学习的综合能源市场交易优化决策[J]．电力系统自动化，2021，45(16)：124-133.
SUN Qingkai，WANG Xiaojun，WANG Yi，et al．Optimal trading decision-making for integrated energy market based on multi-agent nash-Q reinforcement learning[J]．Automation of Electric Power Systems，2021，45(16)：124-133(in Chinese).

[129] YAN Ziming，XU Yan．Real-time optimal power flow：A Lagrangian based deep reinforcement learning approach[J]．IEEE Transactions on Power Systems，2020，35(4)：3270-3273.

[130] SHI Yuanyuan，QU Guannan，LOW S，et al．Stability constrained reinforcement learning for real-time voltage control[C]//2022 American Control Conference(ACC)．Atlanta：IEEE，2022：2715-2721.

[131] 张沛，朱驻军，谢桦．基于深度强化学习近端策略优化的电网无功优化方法[J]．电网技术，2023，47(2)：562-572.
ZHANG Pei，ZHU Zhujun，XIE Hua．Reactive power optimization method based on proximal policy optimization of deep reinforcement learning[J]．Power System Technology，2023，47(2)：562-572(in Chinese).

[132] JIANG Yu，JIANG Zhongping．Robust adaptive dynamic programming for large-scale systems with an application to multimachine power systems[J]．IEEE Transactions on Circuits and Systems II：Express Briefs，2012，59(10)：693-697.

[133] GUO Wentao，LIU Feng，SI J，et al．Approximate dynamic programming based supplementary reactive power control for DFIG wind farm to enhance power system stability[J]．Neurocomputing，2015，170：417-427.

[134] DUAN Jiajun，SHI Di，DIAO Ruisheng，et al．Deep-reinforcement-learning-based autonomous voltage control for power grid operations[J]．IEEE Transactions

on Power Systems，2020，35(1)：814-817.

[135] 窦飞，蔡晖，郭朝辉，等．基于深度强化学习的多端背靠背柔性直流系统直流电压控制[J]．电力系统自动化，2021，45(19)：155-162.
DOU Fei，CAI Hui，GUO Zhaohui，et al．DC voltage control of back-to-back multi-terminal VSC-HVDC system based on deep reinforcement learning[J]．Automation of Electric Power Systems，2021，45(19)：155-162(in Chinese).

[136] 赵冬梅，陶然，马泰屹，等．基于多智能体深度确定策略梯度算法的有功-无功协调调度模型[J]．电工技术学报，2021，36(9)：1914-1925.
ZHAO Dongmei，TAO Ran，MA Taiyi，et al．Active and reactive power coordinated dispatching based on multi-agent deep deterministic policy gradient algorithm[J]．Transactions of China Electrotechnical Society，2021，36(9)：1914-1925(in Chinese).

[137] LIANG Yanchang，GUO Chunlin，DING Zhaohao，et al．Agent-based modeling in electricity market using deep deterministic policy gradient algorithm[J]．IEEE Transactions on Power Systems，2020，35(6)：4180-4192.

[138] FU J，KUMAR A，SOH M，et al. Diagnosing bottlenecks in deep Q-learning algorithms[C]//Proceedings of the 36th International Conference on Machine Learning．Long Beach：PMLR，2019：2021-2030.

[139] ZHANG Shangtong，YAO Hengshuai，WHITESON S．Breaking the deadly triad with a target network[C]//Proceedings of the 38th International Conference on Machine Learning．Online：PMLR，2021：12621-12631.

[140] SINGH V P，KISHOR N，SAMUEL P．Distributed multi-agent system-based load frequency control for multi-area power system in smart grid[J]．IEEE Transactions on Industrial Electronics，2017，64(6)：5151-5160.

[141] 陈艺璇，张孝顺，郭乐欣，等．基于多智能体迁移强化学习算法的电力系统最优碳－能复合流求解[J]．高电压技术，2019，45(3)：863-872.
CHEN Yixuan，ZHANG Xiaoshun，GUO Lexin，et al．Optimal carbon-energy combined flow in power system based on multi-agent transfer reinforcement learning[J]．High Voltage Engineering，2019，45(3)：863-872(in Chinese).

[142] DAI Pengcheng，YU Wenwu，WEN Guanghui，et al．Distributed reinforcement learning algorithm for dynamic economic dispatch with unknown generation cost functions[J]．IEEE Transactions on Industrial Informatics，2020，16(4)：2258-2267.

[143] 胥鹏，王蓓蓓，包宇庆，等．基于网络拓扑资源的配电网在线电压控制方法及其迁移强化学习求解[J]．中国电机工程学报，2020，40(22)：7317-7328.
XU Peng，WANG Beibei，BAO Yuqing，et al．Online voltage control method based on topology resource of network and the solving method via transfer reinforcement learning[J]．Proceedings of the CSEE，2020，40(22)：7317-7328(in Chinese).

[144] WANG Jianhong，XU Wangkun，GU Yunjie，et al．Multi-agent reinforcement learning for active voltage control on power distribution networks[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems．Online：Curran Associates，Inc.，2021：3271-3284.

[145] CANESE L，CARDARILLI G C，DI NUNZIO L，et al．Multi-agent reinforcement learning：A review of challenges and applications[J]．Applied Sciences，2021，11(11)：4948.

[146] DU Wei，DING Shifei．A survey on multi-agent deep reinforcement learning：from the perspective of challenges and applications[J]．Artificial Intelligence Review，2021，54(5)：3215-3238.

[147] LI Yingying，TANG Yujie，ZHANG Runyu，et al．Distributed reinforcement learning for decentralized linear quadratic control：A derivative-free policy optimization approach[J]．IEEE Transactions on Automatic Control，2022，67(12)：6429-6444.

[148] QU Guannan，WIERMAN A，LI Na．Scalable reinforcement learning for multiagent networked systems [J]．Operations Research，2022，70(6)：3601-3628.

[149] ZHAO Wenshuai，QUERALTA J P，WESTERLUND T．Sim-to-real transfer in deep reinforcement learning for robotics：a survey[C]//2020 IEEE Symposium Series on Computational Intelligence(SSCI)．Canberra：IEEE，2020：737-744.

[150] ALI K H，SIGALO M，DAS S，et al．Reinforcement learning for energy-storage systems in grid-connected microgrids：An investigation of online vs．offline implementation[J]．Energies，2021，14(18)：5688.

[151] WANG Tianjing，TANG Yong．Transfer-reinforcement-learning-based rescheduling of differential power grids considering security constraints[J]．Applied Energy，2022，306：118121.

[152] 刘潇，刘书洋，庄韫恺，等．强化学习可解释性基础问题探索和方法综述[J]．软件学报，2023，34(5)：2300-2316.
LIU Xiao，LIU Shuyang，ZHUANG Yunkai，et al．Explainable reinforcement learning：Basic problems exploration and method survey[J]．Journal of Software，2023，34(5)：2300-2316(in Chinese).

[153] HEUILLET A，COUTHOUIS F，DÍAZ-RODRÍGUEZ

N. Explainability in deep reinforcement learning[J]. Knowledge-Based Systems，2021，214：106685.

[154] MACHLEV R，HEISTRENE L，PERL M，et al. Explainable artificial intelligence(XAI) techniques for energy and power systems：Review，challenges and opportunities[J]. Energy and AI，2022，9：100169.

[155] 蒲天骄，乔骥，赵紫璇，等. 面向电力系统智能分析的机器学习可解释性方法研究(一)：基本概念与框架[J/OL]. 中国电机工程学报，2022：1-20[2022-12-08]. https://doi.org/10.13334/j.0258-8013.pcsee.221366.
PU Tianjiao，QIAO Ji，ZHAO Zixuan，et al. Research on interpretable methods of machine learning applied in intelligent analysis of power system(Part I)：basic concept and framework[J/OL]. Proceedings of the CSEE，2022：1-20[2022-12-08]. https://doi.org/10.13334/j.0258-8013.pcsee.221366(in Chinese).

[156] ZHANG Ke，ZHANG Jun，XU Peidong，et al. Explainable AI in deep reinforcement learning models for power system emergency control[J]. IEEE Transactions on Computational Social Systems，2022，9(2)：419-427.

[157] 戴宇欣，陈琪美，高天露，等. 基于加权倾斜决策树的电力系统深度强化学习控制策略提取[J]. 电力信息与通信技术，2021，19(11)：17-23.
DAI Yuxin，CHEN Qimei，GAO Tianlu，et al. Deep reinforcement learning control policy extraction based on weighted oblique decision tree[J]. Electric Power Information and Communication Technology，2021，19(11)：17-23(in Chinese).

[158] BOTVINICK M，RITTER S，WANG J X，et al. Reinforcement learning，fast and slow[J]. Trends in Cognitive Sciences，2019，23(5)：408-422.

[159] DIAO Ruisheng，WANG Zhiwei，SHI Di，et al. Autonomous voltage control for grid operation using deep reinforcement learning[C]//2019 IEEE Power & Energy Society General Meeting(PESGM). Atlanta：IEEE，2019：1-5.

[160] 张峻伟，吕帅，张正昊，等. 基于样本效率优化的深度强化学习方法综述[J]. 软件学报，2022，33(11)：4217-4238.
ZHANG Junwei，LV Shuai，ZHANG Zhenghao，et al. Survey on deep reinforcement learning methods based on sample efficiency optimization[J]. Journal of Software，2022，33(11)：4217-4238(in Chinese).

[161] YU Yang. Towards sample efficient reinforcement learning[C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm，Sweden：AAAI Press，2018：5739-5743.

[162] NACHUM O，GU Shixiang，LEE H，et al. Data-efficient hierarchical reinforcement learning[C]//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal：Curran Associates，Inc.，2018.

[163] LIU Hao，SOCHER R，XIONG Caiming. Taming MAML：Efficient unbiased meta-reinforcement learning [C]//Proceedings of the 36th International Conference on Machine Learning. Long Beach：PMLR，2019：4061-4071.

[164] YARATS D，ZHANG A，KOSTRIKOV I，et al. Improving sample efficiency in model-free reinforcement learning from images[J]. Proceedings of the AAAI Conference on Artificial Intelligence，2021，35(12)：10674-10681.

[165] JIN Chi，YANG Zhuoran，WANG Zhaoran，et al. Provably efficient reinforcement learning with linear function approximation[C]//Proceedings of Thirty Third Conference on Learning Theory. Graz：PMLR，2020：2137-2143.

[166] KARNIADAKIS G E，KEVREKIDIS I G，LU Lu，et al. Physics-informed machine learning[J]. Nature Reviews Physics，2021，3(6)：422-440.

[167] AANGENENT W，KOSTIC D，DE JAGER B，et al. Data-based optimal control[C]//Proceedings of the 2005，American Control Conference，2005. Portland：IEEE，2005：1460-1465.

[168] LEWIS F L，VAMVOUDAKIS K G. Reinforcement learning for partially observable dynamic processes：Adaptive dynamic programming using measured output data[J]. IEEE Transactions on Systems，Man，and Cybernetics，Part B(Cybernetics)，2011，41(1)：14-25.

[169] ZHU Pengfei，LI Xin，POUPART P，et al. On improving deep reinforcement learning for POMDPs[J]. arXiv：1704.07978，2018.

[170] MENG Lingheng，GORBET R，KULIĆ D. Memory-based deep reinforcement learning for POMDPs[C]// 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems(IROS). Prague：IEEE，2021：5619-5626.

[171]EFRONI Y，JIN Chi，KRISHNAMURTHY A，et al. Provable reinforcement learning with a short-term memory[C]//Proceedings of the 39th International Conference on Machine Learning. Baltimore：PMLR，2022.

[172] GARCÍA J，FERNÁNDEZ F. A comprehensive survey on safe reinforcement learning[J]. The Journal of Machine Learning Research，2015，16(1)：1437-1480.

[173] FULTON N，PLATZER A. Safe reinforcement learning via formal methods：Toward safe control through proof

and learning[C]//Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. New Orleans: AAAI Press, 2018.

[174] GROS S, ZANON M, BEMPORAD A. Safe reinforcement learning via projection on a safe set: how to achieve optimality?[J]. IFAC-PapersOnLine, 2020, 53(2): 8076-8081.

[175] CHEN Jinglin, JIANG Nan. Information-theoretic considerations in batch reinforcement learning[C]// Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR, 2019: 1042-1051.

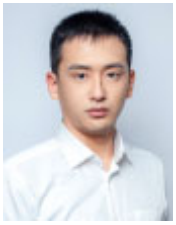
[176] LI Jinning, TANG Chen, TOMIZUKA M, et al. Dealing with the unknown: pessimistic offline reinforcement learning[C]//Proceedings of the 5th Conference on Robot Learning. London: PMLR, 2022: 1455-1464.

[177] LEE S, SEO Y, LEE K, et al. Offline-to-online reinforcement learning via balanced replay and pessimistic Q-ensemble[C]//Proceedings of the 5th Conference on Robot Learning. London: PMLR, 2022: 1702-1712.

[178] CHENG Lefeng, YU Tao. A new generation of AI: A review and perspective on machine learning technologies applied to smart energy and electric power systems[J]. International Journal of Energy Research, 2019, 43(6): 1928-1973.

[179] IBRAHIM M S, DONG Wei, YANG Qiang. Machine learning driven smart electric power systems: Current trends and new perspectives[J]. Applied Energy, 2020, 272: 115237.

[180] YIN Linfei, ZHANG Bin. Time series generative adversarial network controller for long-term smart generation control of microgrids[J]. Applied Energy, 2021, 281: 116069.



毕聪博

在线出版日期：2023-02-23。

收稿日期：2022-12-22。

作者简介：

毕聪博(1997)，男，博士研究生，研究方向为电力系统运行与控制，Thueeq_bcb@outlook.com；

唐聿劼(1991)，男，助理教授，研究方向为数据驱动的基于学习的控制，yujietang@pku.edu.cn；

罗永红(1996)，女，博士研究生，研究方向为数据驱动的电力系统稳定态势感知与控制，13167955468@163.com；

*通信作者：陆超(1977)，男，教授，研究方向为电力系统分析与稳定控制，luchao@tsinghua.edu.cn。

(责任编辑　乔宝榆)