第47卷 第17期 2023年9月10日
DOI：10.7500/AEPS20220228002
电力系统自动化
Automation of Electric Power Systems
Vol. 47 No. 17 Sept. 10，2023

基于深度强化学习的新型电力系统调度优化方法综述

冯 斌[1]，胡轶婕[1]，黄 刚[2]，姜 威[1]，徐华廷[1]，郭创新[1]
（1. 浙江大学电气工程学院，浙江省杭州市 310027；2. 之江实验室，浙江省杭州市 311121）

摘要：随着新能源并网规模不断扩大，能源形式更加灵活多变，电力系统调度运行面临新的挑战。随着系统复杂度和不确定性增加，传统基于物理模型的优化方法难以建立精确的模型进行实时快速求解，而深度强化学习（DRL）可以从历史经验中自适应地学习调度策略并实时决策，避免了复杂的建模过程，以数据驱动的方式应对更高的不确定性和复杂度。文中首先介绍了新型电力系统调度运行问题；然后，介绍了DRL原理及其分类算法；接着，分析了各类DRL算法在求解新型电力系统调度决策问题时的优势与劣势；最后，对需进一步研究的方向进行了展望。

关键词：深度强化学习；新型电力系统；经济调度；最优潮流；机组组合

0 引言

新型电力系统是以确保能源电力安全为基本前提，以绿电消费为主要目标，以坚强智能电网为枢纽平台，以源网荷储互动及多能互补为支撑，具有绿色低碳、安全可控、智慧灵活、开放互动、数字赋能、经济高效基本特征的电力系统[1]。随着“碳达峰·碳中和”目标的提出，新能源在电力能源供给中的占比逐渐增加，将形成新能源占比逐渐提高的新型电力系统[2]。未来，电力占终端能源形式的比例需提高至80%[3]，非化石能源在生产侧的占比要达到80%，光伏、风电等清洁能源装机容量势必逐年增长。新能源的广泛接入与迅速发展使得新型电力系统的随机性、不确定性显著增加，这给传统的调度优化方法带来了极大的挑战。

强化学习（reinforcement learning，RL）拥有强大的自主搜索和学习能力，与监督学习、无监督学习并称现今3种机器学习范式[4]，其侧重于学习实现目标的最优策略。而深度学习（deep learning，DL）[5]通过多层的网络结构，可以对高维数据特征进行抽取，更侧重于对事物的特征提取与感知理解。结合RL与DL的深度强化学习（deep reinforcement learning，DRL）在适应复杂状态环境的同时，能够无需依赖于预测数据即可实现在线实时的调度控制，目前已经在游戏[6]、围棋[7]、机器人控制[8]、城市智慧交通[9]、ChatGPT智能对话等领域得到了广泛应用，在很多场景下甚至能够超越人类表现。

DRL起源于动态规划，其实质是解决一个动态优化问题，理论源于动态规划与马尔可夫决策过程（Markov decision process，MDP），相较于启发式搜索算法更具备理论基础。DRL作为一种数据驱动方法，能够从历史经验中学习决策调度方法，针对非线性、非凸问题具有很好的自适应学习决策能力。目前，大多通过无模型的算法处理，避免了对不确定实时变化的物理模型进行建模，适用于复杂多变的场景。相较于其他传统优化方法，DRL对同一问题模型的不同数据具有更好的泛化能力，以及在相似问题之间具有更好的迁移性，并已在电网频率控制[10]、电压控制[11]等领域得到应用。

本文从DRL原理出发，对DRL算法在新型电力系统调度中的应用现状进行了总结。

1 新型电力系统调度问题

随着新能源接入比例的提高、电网规模的不断扩大，为提高系统整体运行的经济性与可靠性，应协调调度电网的发电资源与用电资源。新型电力系统中的调度问题是为了解决电力系统供需平衡的高维、不确定性强的优化问题。其中，电力系统经济调度（economic dispatch，ED）、最优潮流（optimal power flow，OPF）和机组组合（unit commitment，UC）问题是电力系统运行中的3个关键问题。

1）经济调度问题是以最小化电力系统的总运营成本为目标、满足电力需求和各种运行约束的优化

收稿日期：2022-02-28；修回日期：2022-04-28。
上网日期：2023-04-03。
国家自然科学基金资助项目（U22B2098）；浙江省自然科学基金资助项目（LQ20E070002）；国家留学基金资助项目（202106320157）。

问题。传统的经济调度问题是在满足功率平衡和机组功率边界的前提下,确定各火电发电机组的有功出力,使得总燃料耗量(发电成本)最小。随着新能源出力不确定性的增加,系统的约束条件更加复杂、不确定性更强。

2)最优潮流问题[12]是指在满足电力系统潮流等式约束,以及节点电压、线路潮流、发电机爬坡等不等式约束的情况下,在主网中实现发电成本最小或在配电网中实现网损最小的优化问题。最优潮流与经济调度问题的区别主要在于是否考虑电力系统潮流等式约束。新型电力系统所含风电、光伏等间歇性新能源使得电力系统最优潮流问题,尤其是交流最优潮流问题[13]的求解更加复杂。

3)机组组合问题是在满足系统负荷需求和其他约束条件时实现系统运行成本最小的机组启停计划优化问题。随着大量新能源接入,机组组合方案繁多,不确定性增加,求解更加困难。

传统的优化调度方法往往需要对系统做出一系列假设,同时也难以应对系统动态变化的挑战。随机优化、鲁棒优化、分布式鲁棒优化、启发式优化算法等传统优化算法被用于解决新型电力系统的不确定性问题,但它们都依赖于精准的预测,难以应对新能源出力与负荷需求多变的场景。随机优化常通过采样、机会约束生成等方式将不确定性问题转化为确定性问题,但是算法复杂度随着场景的增加而增加;鲁棒优化通过给出不确定集的方式解决不确定性问题,但是通常其给出的优化结果仅面向最恶劣的场景,过于保守;启发式优化算法,如遗传算法、粒子群算法等,容易陷入局部最优,而且动作复杂度的增加给启发式的优化算法带来严重的维数灾问题,难以稳定收敛。

DRL因其实时决策、不断反馈修正的特性,能够更好地应对新型电力系统新能源的不确定性,可为新型电力系统调度问题提供新的解决途径。

2　DRL原理

2.1　从RL到DRL

RL借鉴了行为主义心理学,是一类特殊的机器学习算法。与监督学习和无监督学习的回归分类目标不同的是,RL是一种最大化未来奖励的决策学习模型,通过与环境交互建立的MDP[14]解决复杂的序列决策问题。RL中常见的概念包括智能体、环境、状态(state,S)、动作(action,A)、奖励(reward,R)。如图1所示,智能体处在环境中,执行动作后获得一定的奖励,而环境由于智能体执行的动作发生状态的变化。依据每一步获得的奖励,通过特定的算法最大化未来的累计奖励是RL算法的核心。详细RL原理见附录A。

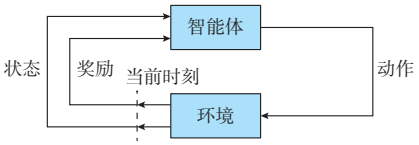

图1　智能体与环境的交互过程

Fig. 1　Interaction process between agent and environment

在传统的RL[15]中,一般可以通过迭代求解贝尔曼最优方程获得最优动作价值函数与状态价值函数,进而指导智能体做出选择。但是在实际场景下,存在着迭代效率低、计算代价大等问题。为此,通常采用参数化的神经网络来近似估计最优动作价值函数和状态价值函数,这也就形成了DRL。

2.2　DRL算法

依据是否有模型,将DRL算法分为基于模型的DRL和无模型的DRL。其中,基于模型的DRL是指智能体可以学习到环境动态变化的参数。在无模型的DRL中,依据智能体的动作选择方式,又可分为基于价值、基于策略、执行者-评论者的算法,其中,执行者-评论者算法也可以看做是结合了基于价值与基于策略的算法。

2.2.1　基于模型的DRL算法

基于模型的DRL算法需要对环境进行建模,然后,基于模型给出策略选择或者动作规划,因而其采样效率较高。该环境通常指状态转移模型,即真实环境的动态变化模型。

结合无模型微调的基于模型的RL[16](model-based RL with model-free fine-tuning,MBMF)是一种基于学习到的环境进行模型预测控制的算法。MBMF首先基于数据集训练神经网络动态模型去学习环境;然后,针对该动态模型执行模型预测控制,并将控制器产生的运行结果进一步添加到神经网络动态模型中进行训练。重复整个迭代训练过程,直至MBMF达到所需的性能表现。

AlphaZero[17]是一种利用已有环境的基于模型的DRL算法。它是AlphaGo[7]的改进,可实现从围棋到各类棋类游戏的智能博弈,通过自主学习环境规划搜索策略。AlphaZero与MuZero[18]通过蒙特卡洛树搜索(Monte Carlo tree search,MCTS)[19]对所学习得到的策略函数进行搜索,实现了动作的多样性探索。

2.2.2　基于价值的DRL算法

基于价值的DRL算法是通过迭代或者训练得到最优动作价值函数,智能体依据最优动作价值函

数选择获得最大的最优动作价值函数所对应的动作，从而实现了策略选择。常见的基于价值的DRL算法包括深度Q学习（deep Q-learning，DQN）[6,20]及其改进算法、优先经验回放[21]、Double Q-learning[22]、Dueling DQN[23]和值分布RL算法中的C51[24]以及Rainbow DQN[25]等。

最早提出的RL算法是基于价值的Q学习[15]与状态-动作-奖励-状态-动作（state-action-reward-state-action，SARSA）[26]算法，它们是通过采用最优贝尔曼方程更新Q值表的方式，迭代得到最优动作价值。

随后，文献[6,20]将卷积神经网络（convolution neural network，CNN）与传统RL算法中的Q学习算法结合，提出了DQN模型。为避免蒙特卡洛更新带来的巨大方差问题，DQN采用时间差分算法更新最优动作价值函数，更新目标如式（1）所示。

$$y_t = r_t + \gamma \max_{a_t} Q(\boldsymbol{s}_{t+1}, \boldsymbol{a}_t; \boldsymbol{w}_t) \tag{1}$$

式中：y_t为t时刻由时间差分算法得到的目标动作价值；r_t为动作得到的奖励；$\gamma \in [0,1]$为奖励衰减因子；$Q(\boldsymbol{s}_{t+1}, \boldsymbol{a}_t; \boldsymbol{w}_t)$为动作价值的神经网络函数；$\boldsymbol{s}_{t+1}$为$t+1$时刻的状态；$\boldsymbol{a}_t$为$t$时刻的动作；$\boldsymbol{w}_t$为$t+1$时刻神经网络参数。

随后，为解决DQN过高估计最优动作价值函数的问题，在Double DQN[22]中引入目标网络，在Dueling DQN[23]中采用竞争架构分别估计优势函数和状态价值函数。采用差异化的优先经验回放[21]提高训练效率，添加高斯噪声以提高动作的探索能力[27]。为充分利用动作价值函数的分布信息，进一步提出了分布式价值的C51算法[24]以及学习分布分位数值的分位数回归深度Q学习（quantile regression DQN，QR-DQN）算法[28]，以及结合上述所有改进的Rainbow DQN[25]算法。

虽然Rainbow DQN算法在离散动作空间的游戏策略问题上取得了不错的效果，但是只能针对离散动作空间进行建模。对于实际问题中常见的连续动作空间则需要进行离散化处理，可能会造成一定动作空间的损失和维数增多的问题。

2.2.3　基于策略的DRL算法

基于策略的DRL算法也可称作是基于策略梯度的DRL，相较于基于价值的DRL，其策略函数可以直接映射到连续动作空间，对于连续控制问题具有更好的效果。

基于策略的DRL是通过最大化奖励较高动作的出现概率，实现未来期望奖励的最大化。这是一种端到端的学习方式，直接优化策略的期望奖励。常见的基于策略的RL算法有：经典的策略梯度RL算法[29]、置信域策略优化（trust region policy optimization，TRPO）[30]算法、近端策略优化（proximal policy optimization，PPO）[31]算法等。

在基于策略的DRL中，采用参数为$\boldsymbol{\theta}$的神经网络来代替策略函数。策略梯度表示形式如式（2）所示。

$$g = (R - b)\nabla_{\boldsymbol{\theta}} \sum_{t=0}^{T-1} \log \pi(\boldsymbol{a}_t | \boldsymbol{s}_t; \boldsymbol{\theta}) \tag{2}$$

式中：g为策略梯度值；R为奖励；b为不依赖于动作的基线；$\boldsymbol{s}_t$为t时刻的状态；T为该情节所经历的时间步；$\pi(\boldsymbol{a}_t|\boldsymbol{s}_t; \boldsymbol{\theta})$为策略函数。梯度项$\nabla_{\boldsymbol{\theta}} \sum_{t=0}^{T-1} \log \pi(\boldsymbol{a}_t|\boldsymbol{s}_t; \boldsymbol{\theta})$为希望将情节获得的奖励向上提高的梯度。

参数更新时将在现有参数$\boldsymbol{\theta}$上加上αg，实现梯度上升，其中，α为学习率。上述训练过程将最大化较高奖励动作的出现概率。

RL算法[29]使用蒙特卡洛方法更新策略梯度，具有较好的稳定性，但是采样效率较低，会带来较大的估计方差。为此在策略学习中减去基线，可有效减少方差。由于基于策略的RL对步长十分敏感，上述方法难以直接选择合适的步长，如果新旧策略差异过大则不利于学习。TRPO[30]通过约束限制新旧策略动作的KL（Kullback-Leibler）散度，避免了策略发生过大参数更新步的情况，解决了策略梯度更新步长的问题。而PPO[31]则通过模型自适应地调整新旧策略动作的KL散度，以保证策略梯度的稳定更新。但是TRPO和PPO都是采用同步更新策略的算法，其每次更新都需要采样大量样本，算法复杂度高、训练效率低，并且其应用也需要大量算力支撑。

2.2.4　执行者-评论者DRL算法

执行者-评论者DRL算法中的执行者算法类似于基于策略的DRL算法，评论者算法类似于基于价值的DRL算法。因此，执行者-评论者DRL算法同时学习策略和价值函数，其框架图如图2所示。执行者-评论者也可以被认为是一种基于策略的DRL算法，特殊之处在于它使用了状态价值函数作为式（2）的基线b，减小了方差，即$A_\pi(\boldsymbol{s}_t, \boldsymbol{a}_t) = Q_\pi(\boldsymbol{s}_t, \boldsymbol{a}_t) - V_\pi(\boldsymbol{s}_t)$，其中，$Q_\pi(\boldsymbol{s}_t, \boldsymbol{a}_t)$为动作价值，$V_\pi(\boldsymbol{s}_t)$为状态价值。$A_\pi(\boldsymbol{s}_t, \boldsymbol{a}_t)$也被称为优势函数，若优势函数大于0，则表示该动作优于平均值，是合理的选择。

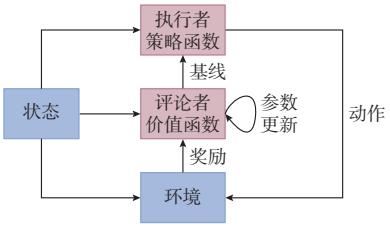

图2　执行者-评论者DRL算法框架
Fig. 2　Framework of actor-critic DRL algorithm

它既结合了基于价值和基于策略DRL算法的优点，也在一定程度上继承了二者的缺点。常见的执行者-评论者DRL算法包括确定性策略梯度(deterministic policy gradient，DPG)算法[32]、深度确定性策略梯度(deep deterministic policy gradient，DDPG)[33]算法、柔性执行者-评论者(soft actor-critic，SAC)[34]算法、异步优势执行者-评论者(asynchronous advantage actor-critic，A3C)[35]算法、双延迟确定性策略梯度(twin delayed deep deterministic policy gradient，TD3)算法[36]等。

DPG每次确定性地探索一个动作，降低了采样需求，能够处理动作空间较大的问题，但为保证未知动作的探索能力，必须采用异步策略更新方法。DDPG在DPG的基础上借鉴了DQN在Q学习基础上改进的思想，利用深度神经网络拟合DDPG中的Q函数，采用异步的Critic估计策略梯度，使训练更加稳定简单。TD3在DDPG的基础上引入了性能更优的Double DQN，并通过取2个Critic之间的最小值避免过拟合，解决了过高估计以及方差过大的问题。过高的估计会使得更新方向与理想情况有偏差，而方差过大会使得训练不稳定。SAC建立在非策略最大熵RL框架[37]上，在实现策略预期回报最大化的同时也具有最大熵，可提升算法的探索能力。

上述异步策略更新算法可以在策略更新时重复利用过去的样本，对样本利用效率高。目前，常见的异步策略更新的DRL算法，均是以DPG为基础的确定性策略算法，如DDPG、TD3等。但是，基于确定性策略的算法对超参数敏感，收敛难度较大。A3C中有多个智能体在中央处理器(central processing unit，CPU)多线程上异步执行，使得样本间的相关性很低。因此，A3C中也没有采用经验回放的机制，而是直接采用同步策略更新机制。

2.2.5　多智能体与分层DRL算法

在DRL的基础上，结合多智能体、分层级等理论，提出了一些适用于更加复杂场景的DRL算法。

1)多智能体DRL算法

考虑到现实复杂的实际环境中，往往不止一个动作发出者，即有许多智能体通过共同交互信息实现合作或竞争，其主要目标是实现共同奖励的最大化与多智能体之间的均衡。早期的多智能体RL，考虑多智能体之间的互相博弈提出了Nash-Q学习算法[38]，这类算法需要大量的存储空间存储Q值，适用于规模较小的问题。

近年来，随着DDPG、A3C等算法拥有更优的性能表现，目前，多智能体DRL大多基于执行者-评论者算法框架，其中，最具有代表性的是多智能体深度确定性策略梯度(multi-agent deep deterministic policy gradient，MADDPG)[39]和反事实基线的多智能体执行者-评论者[40]。它们均采用集中式训练、分布式执行的算法模式，利用所有状态信息集中训练出评论者，每个智能体仅采用自身观测到的信息，执行各自的动作。在智能体动作执行期间，解决了多智能体间信息及时共享的问题。在新型电力系统调度问题中，常见的多区域电网、微电网(microgrid，MG)、综合能源系统都可以采用多智能体DRL算法进行求解。

此外，在基于价值分解的多智能体DRL算法中，多个智能体通过简单加和局部价值函数[41]或采用非线性混合网络[42]联合价值函数，将各主体观测到的局部价值函数合并为联合价值函数。因此，此类算法大多用于共同合作问题。

2)分层DRL算法

一个复杂问题往往会有庞大的状态空间与动作空间，导致实际奖励是非常稀疏的，而分层DRL算法的提出将改善奖励反馈稀疏的问题。分层DRL[43]可以在一些复杂的DRL任务环境下，将最终任务转变为多个子任务的形式，实现DRL任务的分解。通过各子主体策略来形成有效的全局策略。

经典分层强化学习方法是将复杂问题建模为半马尔可夫过程，底层策略建模为MDP问题。经典的分层强化学习算法包括Option[44]、分层抽象机(hierarchies of abstract machines，HAMs)[45]、MAXQ[46]算法等。当今，结合深度学习的分层DRL算法采用2层结构：上层结构每隔一段时间进行调用，根据调用时观测到的状态，给出下层子任务；下层结构作为底层结构，根据当前目标状态和子任务产生动作。例如，分层DQN[47]的双层均采用DQN网络，上层制定一个下层能够实现的小目标并由下层网络实现，待小目标实现后或达到指定时间后，重复指定新的小目标；子策略共享分层DRL算法[48]将子策略参数共享，以提升子任务的训练效率。文献[49]将分层DRL算法应用于多微电网经济调度模型，实现了长短期利益结合的分布式经济调度。

3　DRL在新型电力系统调度中的应用分析

将DRL应用于新型电力系统调度问题时，需要定义DRL中的智能体、环境、状态、动作以及奖励。智能体指动作的发出者，也可认为是系统运行人员；环境指电力系统；状态指环境中各个设备当前的运行状态，如发电机上一时刻出力、电热功率需求、风光实时功率、目前所处的时段等；动作指系统中可以人为控制调节的变量，如发电机出力、储能等；奖励通常是需要实现的目标，如最小化系统运行成本、最大化新能源消纳、最小化电压频率偏差等。关于DRL应用于新型电力系统调度的文献详见附录B。

3.1　经济调度问题

在经济调度问题中需要决策的变量均为连续变量。因此，常采用DDPG、A3C、PPO等具有连续动作空间的DRL算法。

1）大电网

针对含有风光储的大电网经济调度问题，文献[50]在考虑备用的情况下，采用DDPG应对风光荷不确定性以实现系统的动态经济调度，但DDPG不能够实现异步采样。文献[51]依据电网调度运行指令下发的实际特点，考虑联络线功率、风电场出力，采用A3C算法实现多场景并行学习的智能经济调度。

当涉及多区域电网经济调度问题时，由于模型复杂，涉及动作空间大，常采用多智能体的算法降低动作空间复杂度。文献[52]提出的基于通信网络架构（CommNet）的分布式多智能体DRL算法，在训练过程中可使各区域智能体间无须共享光伏、负荷预测数据和设备参数等信息。为避免有效决策信息的损失，文献[53-54]没有利用预测信息，直接采用端到端决策来进一步提升调度的经济性。

2）微电网

针对含有风光储的微电网经济调度问题，文献[55-59]的动作对象均为储能充放电，实现的目标分别为光储充电站收益最大化、微电网经济稳定运行、负荷需求与发电功率的精准匹配、最小化运行成本（并网）和尽量满足负荷需求（孤岛）。文献[59-60]都考虑能源出力的随机性，构建了运行期望最小化奖励函数。考虑到多微电网的动作空间维度以及学习复杂度，需要采用分层分布式的方式实现在线经济调度[49]。

3）虚拟电厂

针对含有风光储的虚拟电厂（virtual power plant，VPP）经济调度问题，文献[61]将工业用户中的可控负荷作为一种调度资源，考虑了光伏、风电、微型燃气轮机的环保与经济成本，基于A3C算法的三层边缘计算框架实现经济运行策略的高效求解。文献[62]考虑了储能系统，基于对抗生成网络生成的场景数据集以及DDPG算法实现虚拟电厂的鲁棒经济调度。但上述文献并未考虑响应信号在虚拟电厂内部的分解，文献[63]则考虑了上级总的响应信号分解问题，并采用锐度感知最小化算法[64]，提升了算法对环境和奖励的鲁棒性。

4）综合能源系统

在含有热、电、天然气等综合能源系统（integrated energy system，IES）经济调度问题中，文献[65]采用DDPG算法使综合能源系统中的热电联供机组的电功率、燃气锅炉输出的热功率、储能的充放电功率的经济调度动作空间处于连续状态。由于DDPG对超参数敏感且动作空间探索不足，采样效率较低，文献[66]采用SAC算法，解决了电-气综合能源系统中天然气系统利用传统优化方法难以凸化和收敛的问题，可有效应对源荷不确定性，并实现RL智能体模型秒级优化调度决策。

考虑到DRL算法对复杂动作空间探索难度大，文献[67]采用双层RL模型，上层采用RL算法实现电池出力调度，下层采用混合整数线性规划求解综合能源系统经济调度问题，避免了约束作为惩罚项带来的DRL算法复杂度增加问题，提升了模型计算效率。

然而上述方法在保证约束的安全性上仍有一些欠缺，需要采用一些保障安全的算法。文献[68]采用循环神经网络构建新能源预测模型[69]，并引入了安全引导函数来保障策略的安全性，实现了综合能源系统的安全低碳经济运行。

相较于大电网、微电网、虚拟电厂，综合能源系统可以实现多能源利用互补。例如，通过热电联供机组实现电力和热量的同时生产；通过燃气锅炉输出热功率；通过电转气单元将电力转换为气体。随着需要控制的设备种类及参数增多，动作空间也将增加，会导致神经网络的训练收敛速度下降，甚至造成维数灾难。多智能体DRL作为一种有效处理多智能体参与的决策方法，也逐渐在大规模综合能源系统的经济调度问题中得到应用。文献[70]将综合能源系统中的多个利益主体建模为多智能体，文献[71-72]将多综合能源区域（园区）建模为多个主体，而文献[73]将电力系统和热力系统分别建模为2个主体。它们均取得了比单一智能体DRL算法更优的收敛速度和经济效益。同时，通过集中训练分散执行的算法流程，可以解决各利益主体之间数据共享的问题。

3.2 最优潮流问题

文献[74]将传统Q学习算法应用于电力系统最优潮流计算领域,实现电力系统有功、无功、多目标的最优潮流计算。

但是,传统的Q学习采用离散动作,会损失一部分动作空间,为此需要采用基于策略或者执行者-评论者的DRL算法。文献[75]基于CloudPSS仿真云平台,验证了基于DDPG的最优潮流计算的可行性;由于DDPG中的评论者网络难训练、不稳定,文献[76]虽然基于DDPG算法构建了执行者网络,但没有使用评论者网络,而是基于拉格朗日数学解析推导得到了确定性梯度。由于PPO相比于DDPG具有更高的采样效率、更稳定的学习策略,以及更容易调节的超参数,文献[77-78]采用基于模仿学习的PPO算法求解交流最优潮流问题。

前述的最优潮流问题是针对主网的,而配电网由于没有大型发电机组,其研究对象是在满足潮流约束的同时,通过潮流合理分配使得网损最小。文献[79]基于PPO算法控制储能有功功率、无功功率以及风电的无功功率,实现了在不违反电压和电池储能容量约束的情况下配电网网损的最小化。文献[80]采用完全分布式的PPO算法,实现了不平衡配电网的光伏有功功率最大化输出与电压稳定。

针对互联的微电网,由于其动作空间大,传统单一智能体算法难以满足计算需求,需要建模为多智能体DRL问题求解。文献[81]依据智能体的连续离散动作空间,设置了双层DRL,并将潮流等式约束设置在环境中;文献[82]将潮流等安全约束构建成梯度信息,保证最优控制策略产生安全可行的决策方案。

由于并不是所有场景下的调度问题都是非凸的,可以将凸的子问题抽离出来,构建优化问题与DRL结合的双层求解结构。文献[83]将居民微电网的最优运行成本问题建模成混合整数二阶锥的优化问题,并将其转化为MDP主问题与最优潮流二阶锥优化子问题,主问题采用MuZero[18]算法得到较优的在线优化结果。文献[84]针对互联微电网在信息不全情况下的潮流能量管理问题,考虑在配电网层面只能获取公共连接点(point of common coupling,PCC)处的功率信息,设计了双层算法。在上层基于改进的Q学习实现互联微电网购售电成本最优,在下层针对单个微电网实现最优潮流。文献[81]虽然也采用了双层DRL,但实际上是将离散动作空间和连续动作空间作为前后2层DRL的决策空间。

安全约束最优潮流[85]增加了可靠性约束来确保电力系统能够承受一定预想故障的冲击。由于安全约束最优潮流需要搜索预想故障集,如果采用基于优化的交流最优潮流,其计算量也非常大;而DRL方法的提出,将有助于在交流最优潮流的基础上实现安全约束最优潮流。文献[86]以最小化约束越限为奖励,以提升系统在各种随机场景下的$N-1$安全性为核心,采用A3C算法结合电力领域知识在减小负荷削减量的同时降低了系统运行成本。

DRL算法能够在一定程度上解决电力系统交流最优潮流的精确求解问题,尤其是在非凸约束增多时,优化求解复杂度会急剧提升。而DRL在处理类似问题时可以进行精确建模,而不必为实现凸优化而损失模型精度,甚至可以得到比凸松弛后的交流最优潮流优化问题更经济的解。此外,DRL算法在需要大规模搜索时也有一定优势。

由于最优潮流问题需要考虑潮流等式约束,因而相较于经济调度问题,其动作空间受到一定的限制,这也是当前基于DRL算法求解最优潮流的难点。这需要保证在潮流等式约束被满足的同时,处理新能源出力的不确定性并寻求最优发电调度计划。现阶段文献主要将潮流等式约束放在环境中处理,较少文献将潮流等式约束融合至策略产生的约束中,形成安全的策略[82]。

3.3 机组组合问题

文献[87]采用RL算法求解机组组合问题,而文献[88]采用分布式Q学习算法,因仅涉及局部通信,提高了求解的鲁棒性。但是,Q学习算法的动作空间受Q表格的限制,难以处理高维动作状态。为此,文献[89]采用深度神经网络逼近Q函数的DQN算法实现高维机组组合动作空间的探索。为应对新能源出力的不确定性,文献[90]针对随机波动的光伏出力,采用全连接神经网络拟合Q值求解考虑光伏出力的机组组合问题。

由于机组组合的动作空间随着机组数量而急剧增长,在现有文献中,Q学习算法最多仅能应用于含12台机组的算例。为进一步克服机组动作空间随机组数量呈几何增长的问题,文献[91]采用引导树搜索方法实现了对动作空间的快速高效搜索,可求解30台机组组合问题,相比于混合整数线性规划算法,可减少机组的频繁动作,并在降低系统运行成本的同时减少了负荷损失概率。

通常在机组组合问题中,除决策机组启停的离散量外,还需要同时给出机组出力的连续决策变量。文献[89,91]采用Lambda迭代法进行求解;文献[88]将机组组合和经济调度问题建模为一个问题,将连续机组出力作为动作对象,动作空间则满足

机组启停等约束。文献[92]采用SAC确定机组启停计划，然后通过Cplex求解器求解单时段优化问题得到机组出力。而文献[87，90]并未提及机组出力的决策过程。

在机组组合问题中，机组启停动作空间是一个离散的动作空间。采用诸如DQN、PPO等一般的DRL算法难以有效应对机组数增加而带来动作空间维度呈指数增长的问题。因而，基于一般的DRL算法仅能够解决机组数较少的机组组合问题，并且较少涉及新能源接入。但一般的DRL算法对环境的探索能力有限，需要结合树搜索算法或者智能体提前预知一定的环境模型信息，进而提升或引导智能体对高维动作空间的探索效率。

机组组合问题作为一个长时间序列决策问题，即使采用先进的DRL技术也难以实现较好的决策，目前在仿真算例中仍存在较多的问题亟待解决。其中，一个较为关键的问题是用电计划无法完全被满足。在理论研究中，常将用电计划满足程度表述为失负荷风险。由于机组组合的动作空间极大，在机组数量较多、测试时间较长的情况下，失负荷通常是不可避免的。因此，后续的研究重点是改进动作空间的建模形式或采用学习能力更强的算法等以确保用电计划完全满足。

3.4　应用前景分析

由于电网对于安全性和供电可靠性要求较高，实际落地应用不可能一蹴而就。考虑到决策的稳定性、安全性以及误决策的危害，可以先在配电网或用户侧进行一些尝试，然后，再从小区域低电压等级慢慢推广到大区域高电压等级。在配电网侧，由于涉及的设备种类多样、波动性较大，对于算法的实时性要求高，可以采用DRL算法进行实时经济调度、设备出力控制、电压控制等，以实现配电网众多设备的安全实时经济运行。在用户侧，可以实时获取价格信号和屋顶光伏等新能源出力信息，采用DRL算法实时控制需求响应、家用电器、温控负荷等。文献[93]将RL算法应用于美国科罗拉多州一个包含27个家庭的微电网中，应用结果表明，采用RL算法可大幅度降低用户用电成本，实现秒级别的优化控制。文献[94]采用拟合Q迭代算法实现电热水器的控制。该项目是住宅需求响应试点项目的一部分，其中，10台电热水器用于直接负荷控制，每台电热水器配备了8个温度传感器和1个可控功率加热装置。在试点项目中，相比于恒温控制器，采用RL算法可使电热水器的总能耗成本降低15%。2021年5—6月，上海某写字楼中央冷水机组采用RL算法控制冷却机组和冷却水泵来重设定点温度[95]，实现了近似专家系统的控制效果，并验证了RL决策系统的鲁棒性、稳定性和学习速度。

在大电网侧，随着新能源广泛接入，源荷波动愈加剧烈，系统对于日内实时优化的需求上升。可以先采取数据接入、辅助决策方式进行试点运行。如果在试点过程中出现错误，则需要对算法进一步校验，必要时可以增加一些人工调度经验规则，采用数据知识混合驱动的方法保证决策的正确性。常见的实时调度场景包括日前和日内的实时计划动态快速调整、电力市场实时的报价出清策略等。文献[96]所研发的电网脑于2019年11月部署在中国江苏电网调控中心安全Ⅰ区。电网脑能在满足调控需求的前提下，在20 ms内对电压、潮流越界等问题提供解决方案，快速消除风险，同时降低约3.5%的网损。该成果可用于辅助调度员对电压与联络线潮流进行控制，进一步可作为全自动化调度的基础技术手段。

在海量数据场景下，DRL作为一种数据驱动的决策方案，能够在保证目标最优性的同时更快速地求解目标函数，获得比传统方法更高效经济的策略[86，91]。例如，在风险评估中，DRL可以快速搜索高风险级联故障[97-98]，也可以将DRL与电力系统运筹优化方法深度结合，通过DRL加速优化计算或者实现精确建模与求解。

4　研究方向展望

DRL算法能够对智能体进行针对性训练，并能够根据场景的变化快速求得最优管理策略，满足电网运行的实时性要求。但DRL作为一种基于深度神经网络的算法，需要大量学习仿真数据，并且所得到的结果较难解释。电力系统调度是电力系统的核心环节，一般不允许出现差错。若DRL在电力系统调度中获得应用，还需要在以下方面做进一步深入的研究。

1）建立真实的电网仿真环境

DRL需要大量学习仿真数据。在电力系统中，通常需要单独搭建适配于电力系统的环境，智能体在与环境交互的过程中，产生大量情节，这也就是DRL需要学习的仿真数据。DRL的目标是最大化奖励，因此，可以通过奖励的设置对违反的约束给予惩罚，将需要实现的经济性、安全性目标设置在奖励中。考虑到DRL的训练需要搭建类似于Gym[99]的电网环境，当前已有不少开源工作者构建了类似的开源环境库，例如，Gym-ANM[100]、PowerGym[101]、Grid2op[102]等。未来，需要基于数字孪生，搭建电网仿真系统，加强数字资源的积累，为应用提供基础。

2)算法性能的提升

随着建模对象和环境逐渐复杂,在大规模复杂环境下DRL收敛求解时间也会随之增加。如果在实际中求解一个大规模复杂新型电力系统调度问题时,必然会遇到维度灾难问题。当动作空间维数过大时,可搜索的动作空间将很大,进而影响DRL收敛速度和动作的准确性。此外,如果是类似机组组合问题的0-1离散变量过多,也会加剧DRL训练的难度。随着DRL理论的不断发展,未来可以考虑引入模仿学习、元学习的思想[103],以便缩短复杂环境下智能体的培训时间,提高性能。

在与环境交互计算方面,当前智能体与环境的模拟交互过程以及数据的传输通信仍然是通过CPU完成的。如果能够开发类似于Isaac Gym的图形处理器(graphics processing unit,GPU)环境,环境的模拟和神经网络的训练都将置于GPU内,使得数据直接从内存传递到GPU的训练框架中(如PyTorch),不受CPU数据传输限制,则将大大加快目前的训练速度,进一步提高DRL求解大规模复杂问题的性能。

3)安全性研究

由于DRL方法输出的决策存在不确定性,其安全性不如传统优化算法,可能会给出不符合电网安全运行的结果,这时便需要算法有能力给出规避机制,实现电力系统的安全稳定运行。对于新能源全部消纳的要求,可以允许存在一定的弃风弃光,但在有严格物理安全约束要求时,如果DRL不能够完全确保得出的决策满足安全约束,将会导致系统安全问题。当前许多研究基于DRL的调度文献未涉及系统安全约束的问题,即使是涉及系统安全性的文献,也基本是将约束建模成奖励函数惩罚项的形式,极少从数学理论上证明DRL算法可满足安全约束条件。也有将约束在建模过程中直接融合在MDP过程中,形成安全可靠的DRL算法。进一步,也可尝试采用安全RL算法[104]保证策略操作的安全性。

4)可解释性研究

传统基于价值或基于策略的DRL算法,具备强逻辑性和可解释性。但神经网络模型也被称为黑盒子模型,缺乏一定的解释性。而DRL是在RL的基础上,引入了神经网络来拟合价值函数或(和)策略函数,对复杂问题的建模具有更好的实验效果。但是,神经网络的引入不利于其可解释性,难以在实际应用中从原理上说服调度人员依据DRL算法给出的决策进行操作。未来,可结合可解释性机器学习给出可解释性的策略动作,提升DRL的可解释性,让调度人员更易于接受人工智能算法的决策结果。

5)迁移性和鲁棒性研究

目前,研究性论文中智能体所处的环境都是电力系统仿真模拟环境,数据均为理想化的数据,不存在数据干扰的情况。而在实际运行的电力系统环境下,如何保证DRL算法的正确性、保证模型的鲁棒性是值得考虑的问题。文献[63]通过使用锐度感知最小化[64]实现了噪声的鲁棒性,此外,在DRL算法领域也出现了鲁棒DRL算法[105],这也是未来可以尝试的解决方法。

5 结语

本文介绍了新型电力系统调度问题,阐述了基于模型、基于价值、基于策略和执行者-评论者的DRL算法原理,以及在调度中可尝试应用的DRL算法。在经济调度问题中,分别从大电网、微电网、虚拟电厂、综合能源系统角度总结了DRL应用的结果;在最优潮流问题中,以交流最优潮流模型为基础,总结了主网、配电网、微电网以及安全约束最优潮流问题的DRL解决方案;在机组组合问题中,总结了火电发电机组的机组组合和考虑新能源的机组组合问题。最后,分析了当前应用前景,并论述了未来研究方向。

本文受国家自然科学基金项目(52007173,U19B2042)资助,谨此致谢!

附录见本刊网络版(http://www.aeps-info.com/aeps/ch/index.aspx),扫英文摘要后二维码可以阅读网络全文。

参考文献

[1]《新型电力系统技术研究报告》正式发布[EB/OL].[2022-01-14].https://www.tsinghua.edu.cn/info/1175/90380.htm.
Official release of "Research Report on New Power System Technology"[EB/OL]. [2022-01-14]. https://www.tsinghua.edu.cn/info/1175/90380.htm.

[2]舒印彪,陈国平,贺静波,等.构建以新能源为主体的新型电力系统框架研究[J].中国工程科学,2021,23(6):61-69.
SHU Yinbiao, CHEN Guoping, HE Jingbo, et al. Building a new electric power system based on new energy sources[J]. Strategic Study of CAE, 2021, 23(6): 61-69.

[3]加快构建新型电力系统[EB/OL].[2021-11-01]. http://www.xinhuanet.com/comments/2021-08/03/c_1127723605.htm.
Speeding up the construction of new power system [EB/OL]. [2021-11-01]. http://www.xinhuanet.com/comments/2021-08/03/c_1127723605.htm.

[4]SUTTON R S, BARTO A G. Reinforcement learning: an introduction[M]. 2nd ed. Cambridge, USA: A Bradford Book, 2018.

[5]LECUN Y, BENGIO Y, HINTON G. Deep learning[J].

Nature, 2015, 521(7553): 436-444.
[6] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.
[7] 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.
[8] KOBER J, BAGNELL J A, PETERS J. Reinforcement learning in robotics: a survey[J]. International Journal of Robotics Research, 2013, 32(11): 1238-1274.
[9] WEI H, ZHENG G J, YAO H X, et al. IntelliLight: a reinforcement learning approach for intelligent traffic light control[C]// 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, August 19-23, 2018, London, United Kingdom: 2496-2505.
[10] 王力成，邓宝华，黄刚，等.知识-数据混合驱动的电网频率协同控制算法[J].中国电机工程学报，2022，42(23)：8523-8534.
WANG Licheng, DENG Baohua, HUANG Gang, et al. Coordinated system frequency control with a hybrid knowledge-data driven algorithm[J]. Proceedings of the CSEE, 2022, 42(23): 8523-8534.
[11] HOSSAIN R R, HUANG Q H, HUANG R K. Graph convolutional network-based topology embedded deep reinforcement learning for voltage stability control[J]. IEEE Transactions on Power Systems, 2021, 36(5): 4848-4851.
[12] CARPENTIER J. Contribution to the economic dispatch problem[J]. Bulletin de la Societe Francoise des Electriciens, 1962, 3(8): 431-447.
[13] 赵晋泉，叶君玲，邓勇.直流潮流与交流潮流的对比分析[J].电网技术，2012，36(10)：147-152.
ZHAO Jinquan, YE Junling, DENG Yong. Comparative analysis on DC power flow and AC power flow[J]. Power System Technology, 2012, 36(10): 147-152.
[14] BELLMAN R. A Markovian decision process[J]. Indiana University Mathematics Journal, 1957, 6(4): 679-684.
[15] WATKINS C J C H, DAYAN P. Q-learning[J]. Machine Learning, 1992, 8(3): 279-292.
[16] NAGABANDI A, KAHN G, FEARING R S, et al. Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning[C]// 2018 IEEE International Conference on Robotics and Automation (ICRA), May 21-25, 2018, Brisbane, Australia: 7559-7566.
[17] 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.
[18] SCHRITTWIESER J, ANTONOGLOU I, HUBERT T, et al. Mastering Atari, Go, chess and shogi by planning with a learned model[J]. Nature, 2020, 588(7839): 604-609.
[19] COULOM R. Efficient selectivity and backup operators in Monte-Carlo tree search[C]// 5th International Conference on Computers and Games, May 29-31, 2006, Berlin, Germany: 72-83.
[20] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing Atari with deep reinforcement learning[EB/OL]. [2021-12-18]. https://arxiv.org/abs/1312.5602.
[21] SCHAUL T, QUAN J, ANTONOGLOU I, et al. Prioritized experience replay[EB/OL]. [2021-12-25]. https://arxiv.org/abs/1511.05952.
[22] VAN HASSELT H, GUEZ A, SILVER D. Deep reinforcement learning with double Q-learning[C]// 30th AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, USA: 2094-2100.
[23] WANG Z Y, SCHAUL T, HESSEL M, et al. Dueling network architectures for deep reinforcement learning[C]// 33rd International Conference on Machine Learning, June 19-24, 2016, New York, USA: 1995-2003.
[24] BELLEMARE M G, DABNEY W, MUNOS R. A distributional perspective on reinforcement learning[C]// 34th International Conference on Machine Learning, August 6-11, 2017, Sydney, Australia: 449-458.
[25] HESSEL M, MODAYIL J, VAN HASSELT H, et al. Rainbow: combining improvements in deep reinforcement learning[EB/OL]. [2022-01-10]. https://arxiv. org/abs/1710.02298v1.
[26] RUMMERY G, NIRANJAN M. On-line Q-learning using connectionist systems[EB/OL]. [2022-01-10]. https://www.mendeley. com/catalogue/562a4959-6052-3b61-bbde-137a870550a3/.
[27] FORTUNATO M, AZAR M G, PIOT B, et al. Noisy networks for exploration[EB/OL]. [2022-01-10]. https://arxiv.org/abs/1706.10295.
[28] DABNEY W, ROWLAND M, BELLEMARE M, et al. Distributional reinforcement learning with quantile regression[EB/OL]. [2022-01-10]. https://arxiv. org/pdf/1710.10044.pdf.
[29] WILLIAMS R J. Simple statistical gradient-following algorithms for connectionist reinforcement learning[J]. Machine Learning, 1992, 8(3): 229-256.
[30] SCHULMAN J, LEVINE S, MORITZ P, et al. Trust region policy optimization[C]// 32nd International Conference on International Conference on Machine Learning, July 6-11, 2015, Lille, France: 1889-1897.
[31] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[EB/OL]. [2022-01-10]. https://arxiv.org/abs/1707.06347.
[32] SILVER D, LEVER G, HEESS N, et al. Deterministic policy gradient algorithms[EB/OL]. [2022-01-10]. https://dl.acm.org/doi/abs/10.5555/3044805.3044850.
[33] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[EB/OL]. [2022-01-10]. https://arxiv.org/abs/1509.02971.
[34] HAARNOJA T, ZHOU A, ABBEEL P, et al. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor[EB/OL]. [2022-01-10]. https://arxiv.org/abs/1801.01290.
[35] MNIH V, BADIA A P, MIRZA M, et al. Asynchronous methods for deep reinforcement learning[EB/OL]. [2022-01-10]. https://arxiv.org/pdf/1602.01783.pdf.
[36] FUJIMOTO S, VAN HOOF H, MEGER D. Addressing

function approximation error in actor-critic methods[C]// 35th International Conference on Machine Learning (ICML), July 10-15, 2018, Stockholm, Sweden: 1587-1596.

[37] HAARNOJA T, TANG H R, ABBEEL P, et al. Reinforcement learning with deep energy-based policies[C]// 34th International Conference on Machine Learning, August 6-11, 2017, Sydney, Australia: 1352-1361.

[38] HU J, WELLMAN M P. Nash Q-learning for general-sum stochastic games[J]. Journal of Machine Learning Research, 2004, 6(4): 1039-1069.

[39] LOWE R, WU Y, TAMAR A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments[C]// 31st International Conference on Neural Information Processing Systems, December 4-9, 2017, Long Beach, USA: 6382-6393.

[40] FOERSTER J, FARQUHAR G, AFOURAS T, et al. Counterfactual multi-agent policy gradients[C]// 32nd AAAI Conference on Artificial Intelligence, February 2-7, 2018, New Orleans, USA: 2974-2982.

[41] SUNEHAG P, LEVER G, GRUSLYS A, et al. Value-decomposition networks for cooperative multi-agent learning based on team reward[EB/OL]. [2022-01-10]. https://dl.acm.org/doi/pdf/10.5555/3237383.3238080.

[42] RASHID T, SAMVELYAN M, DE WITT C S, et al. QMIX: monotonic value function factorisation for deep multi-agent reinforcement learning[EB/OL]. [2022-02-10]. https://arxiv.org/abs/1803.11485.

[43] BARTO A G, MAHADEVAN S. Recent advances in hierarchical reinforcement learning[J]. Discrete Event Dynamic Systems, 2003, 13(4): 341-379.

[44] SUTTON R S, PRECUP D, SINGH S. Between MDPs and semi-MDPs: a framework for temporal abstraction in reinforcement learning[J]. Artificial Intelligence, 1999, 112(1/2): 181-211.

[45] PARR R, RUSSELL S. Reinforcement learning with hierarchies of machines[EB/OL]. [2022-01-10]. https://dl.acm.org/doi/10.5555/302528.302894.

[46] DIETTERICH T G. Hierarchical reinforcement learning with the MAXQ value function decomposition[J]. Journal of Artificial Intelligence Research, 2000, 13(1): 227-303.

[47] KULKARNI T D, NARASIMHAN K R, SAEEDI A, et al. Hierarchical deep reinforcement learning: integrating temporal abstraction and intrinsic motivation[C]// 30th International Conference on Neural Information Processing Systems, December 5-10, 2016, Barcelona, Spain: 3682-3690.

[48] FLORENSA C, DUAN Y, ABBEEL P. Stochastic neural networks for hierarchical reinforcement learning[EB/OL]. [2022-01-10]. https://arxiv.org/abs/1704.03012.

[49] HAO R, LU T G, AI Q, et al. Distributed online dispatch for microgrids using hierarchical reinforcement learning embedded with operation knowledge[J/OL]. IEEE Transactions on Power Systems[2022-01-10]. https://ieeexplore.ieee.org/document/9464628.

[50] 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.

[51] GUAN J, TANG H, WANG K, et al. A parallel multi-scenario learning method for near-real-time power dispatch optimization[J]. Energy, 2020, 202: 117708.

[52] 张津源,蒲天骄,李烨,等.基于多智能体深度强化学习的分布式电源优化调度策略[J].电网技术,2022,46(9):3496-3504.
ZHANG Jinyuan, PU Tianjiao, LI Ye, et al. Multi-agent deep reinforcement learning based optimal dispatch of distributed generators[J]. Power System Technology, 2022, 46(9): 3496-3504.

[53] 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.

[54] HUANG S Y, LI P, YANG M, et al. A control strategy based on deep reinforcement learning under the combined wind-solar storage system[J]. IEEE Transactions on Industry Applications, 2021, 57(6): 6547-6558.

[55] 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.

[56] HAO R, LU T, AI Q, et al. Distributed online learning and dynamic robust standby dispatch for networked microgrids[J]. Applied Energy, 2020, 274: 115256.

[57] YU Y J, CAI Z F, LIU Y C. Double deep Q-learning coordinated control of hybrid energy storage system in island micro-grid[J]. International Journal of Energy Research, 2021, 45(2): 3315-3326.

[58] BUI V H, HUSSAIN A, KIM H M. Double deep Q-learning-based distributed operation of battery energy storage system considering uncertainties[J]. IEEE Transactions on Smart Grid, 2020, 11(1): 457-469.

[59] 冯昌森,张瑜,文福拴,等.基于深度期望Q网络算法的微电网能量管理策略[J].电力系统自动化,2022,46(3):14-22.
FENG Changsen, ZHANG Yu, WEN Fushuan, et al. Energy management strategy for microgrid based on deep expected Q network algorithm[J]. Automation of Electric Power Systems, 2022, 46(3): 14-22.

[60] 余宏晖,林声宏,朱建全,等.基于深度强化学习的微电网在线优化[J/OL].电测与仪表[2022-01-10].https://kns.cnki.net/kcms2/article/abstract?v=3uoqIhG8C45S0n9fL2suRadTyEVl2pW9UrhTDCdPD64iLFH7p67cuNMs1KaGHuIMxc216qIMqJijn2EpgXJx3VI_73uBFOAb&uniplatform=NZKPT.
YU Honghui, LIN Shenghong, ZHU Jianquan, et al. On-line optimization of microgrid based on deep reinforcement learning[J/OL]. Electrical Measurement & Instrumentation[2022-01-

10]. https://kns.cnki.net/kcms2/article/abstract? v=3uoqIhG8C45S0n9fL2suRadTyEVl2pW9UrhTDCdPD64iLFH7p67cuNMs1KaGHuIMxc216qIMqJijn2EpgXJx3VI_73uBFOAb&uniplatform=NZKPT.

[61] LIN L, GUAN X, PENG Y, et al. Deep reinforcement learning for economic dispatch of virtual power plant in Internet of energy[J]. IEEE Internet of Things Journal, 2020, 7(7): 6288-6301.

[62] FANG D W, GUAN X, HU B R, et al. Deep reinforcement learning for scenario-based robust economic dispatch strategy in Internet of energy[J]. IEEE Internet of Things Journal, 2021, 8(12): 9654-9663.

[63] YI Z K, XU Y L, WANG X, et al. An improved two-stage deep reinforcement learning approach for regulation service disaggregation in a virtual power plant[J]. IEEE Transactions on Smart Grid, 2022, 13(4): 2844-2858.

[64] FORET P, KLEINER A, MOBAHI H, et al. Sharpness-aware minimization for efficiently improving generalization[EB/OL]. [2022-01-10]. https://arxiv.org/abs/2010.01412.

[65] 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.

[66] 乔骥，王新迎，张擎，等.基于柔性行动器-评判器深度强化学习的电-气综合能源系统优化调度[J].中国电机工程学报，2021，41(3)：819-833.
QIAO Ji, WANG Xinying, ZHANG Qing, et al. Optimal dispatch of integrated electricity-gas system with soft actor-critic deep reinforcement learning [J]. Proceedings of the CSEE, 2021, 41(3): 819-833.

[67] 聂欢欢，张家琦，陈颖，等.基于双层强化学习方法的多能园区实时经济调度[J].电网技术，2021，45(4)：1330-1336.
NIE Huanhuan, ZHANG Jiaqi, CHEN Ying, et al. Real-time economic dispatch of community integrated energy system based on a double-layer reinforcement learning method [J]. Power System Technology, 2021, 45(4): 1330-1336.

[68] QIU D, DONG Z, ZHANG X, et al. Safe reinforcement learning for real-time automatic control in a smart energy-hub [J]. Applied Energy, 2022, 309: 118403.

[69] 冯斌，郭亦宗，陈页，等.基于GRU多步预测技术的云储能充放电策略[J].电力系统自动化，2021，45(9)：46-54.
FENG Bin, GUO Yizong, CHEN Ye, et al. Charging and discharging strategy of cloud energy storage based on GRU multi-step prediction technology [J]. Automation of Electric Power Systems, 2021, 45(9): 46-54.

[70] 刘洪，李吉峰，葛少云，等.基于多主体博弈与强化学习的并网型综合能源微网协调调度[J].电力系统自动化，2019，43(1)：40-48.
LIU Hong, LI Jifeng, GE Shaoyun, et al. Coordinated scheduling of grid-connected integrated energy microgrid based on multi-agent game and reinforcement learning[J]. Automation of Electric Power Systems, 2019, 43(1): 40-48.

[71] 杨照，黄少伟，陈颖.基于多智能体强化学习的多园区综合能源系统协同优化运行研究[J].电工电能新技术，2021，40(8)：1-10.
YANG Zhao, HUANG Shaowei, CHEN Ying. Research on cooperative optimal operation of multi-park integrated energy system based on multi agent reinforcement learning [J]. Advanced Technology of Electrical Engineering and Energy, 2021, 40(8): 1-10.

[72] 李昊，刘畅，苗博，等.考虑冷热电互补及储能系统的多园区综合能源系统协调优化调度[J].储能科学与技术，2022，11(5)：1482-1491.
LI Hao, LIU Chang, MIAO Bo, et al. Coordinative optimal dispatch of multi-park integrated energy system considering complementary cooling, heating and power and energy storage systems[J]. Energy Storage Science and Technology, 2022, 11(5): 1482-1491.

[73] 董雷，刘雨，乔骥，等.基于多智能体深度强化学习的电热联合系统优化运行[J].电网技术，2021，45(12)：4729-4738.
DONG Lei, LIU Yu, QIAO Ji, et al. Optimal dispatch of combined heat and power system based on multi-agent deep reinforcement learning[J]. Power System Technology, 2021, 45(12): 4729-4738.

[74] 胡细兵.基于强化学习算法的最优潮流研究[D].广州：华南理工大学，2011.
HU Xibing. The research of optimal power flow based on reinforcement learning[D]. Guangzhou: South China University of Technology, 2011.

[75] NIE H H, CHEN Y, SONG Y K, et al. A general real-time OPF algorithm using DDPG with multiple simulation platforms [C]// 2019 IEEE Innovative Smart Grid Technologies-Asia (ISGT Asia), May 21-24, 2019, Chengdu, China: 3713-3718.

[76] YAN Z M, XU Y. Real-time optimal power flow: a Lagrangian based deep reinforcement learning approach [J]. IEEE Transactions on Power Systems, 2020, 35(4): 3270-3273.

[77] ZHOU Y H, ZHANG B, XU C L, et al. A data-driven method for fast AC optimal power flow solutions via deep reinforcement learning[J]. Journal of Modern Power Systems and Clean Energy, 2020, 8(6): 1128-1139.

[78] ZHOU Y H, LEE W J, DIAO R S, et al. Deep reinforcement learning based real-time AC optimal power flow considering uncertainties[J]. Journal of Modern Power Systems and Clean Energy, 2021, 10(5): 1098-1109.

[79] CAO D, HU W H, XU X, et al. Deep reinforcement learning based approach for optimal power flow of distribution networks embedded with renewable energy and storage devices [J]. Journal of Modern Power Systems and Clean Energy, 2021, 9(5): 1101-1110.

[80] EL HELOU R, KALATHIL D, XIE L. Fully decentralized reinforcement learning-based control of photovoltaics in distribution grids for joint provision of real and reactive power [J]. IEEE Open Access Journal of Power and Energy, 2021, 8: 175-185.

[81] 巨云涛，陈希.基于双层多智能体强化学习的微网群分布式有功无功协调优化调度[J].中国电机工程学报，2022，42(23)：

8534-8548.
JU Yuntao, CHEN Xi. Distributed active and reactive power coordinated optimal scheduling of networked microgrids based on two-layer multi-agent reinforcement learning[J]. Proceedings of the CSEE, 2022, 42(23): 8534-8548.

[82] ZHANG Q Z, DEHGHANPOUR K, WANG Z Y, et al. Multi-agent safe policy learning for power management of networked microgrids[J]. IEEE Transactions on Smart Grid, 2021, 12(2): 1048-1062.

[83] SHUAI H, HE H B. Online scheduling of a residential microgrid via Monte-Carlo tree search and a learned model[J]. IEEE Transactions on Smart Grid, 2021, 12(2): 1073-1087.

[84] ZHANG Q Z, DEHGHANPOUR K, WANG Z Y, et al. A learning-based power management method for networked microgrids under incomplete information [J]. IEEE Transactions on Smart Grid, 2020, 11(2): 1193-1204.

[85] CAPITANESCU F, MARTINEZ RAMOS J L, PANCIATICI P, et al. State-of-the-art, challenges, and future trends in security constrained optimal power flow [J]. Electric Power Systems Research, 2011, 81(8): 1731-1741.

[86] 严梓铭,徐岩.结合深度强化学习与领域知识的电力系统拓扑结构优化[J].电力系统自动化,2022,46(1):60-68.
YAN Ziming, XU Yan. Topology optimization of power systems combining deep reinforcement learning and domain knowledge[J]. Automation of Electric Power Systems, 2022, 46(1): 60-68.

[87] JASMIN E A, IMTHIAS AHAMED T P, JAGTHY RAJ V P. Reinforcement learning solution for unit commitment problem through pursuit method [C]// 2009 International Conference on Advances in Computing, Control, and Telecommunication Technologies, December 28-29, 2009, Bangalore, India: 324-327.

[88] LI F Y, QIN J H, ZHENG W X. Distributed Q-learning-based online optimization algorithm for unit commitment and dispatch in smart grid[J]. IEEE Transactions on Cybernetics, 2020, 50(9): 4146-4156.

[89] 温裕鑫,杨军,朱旭.基于深度强化学习的电网机组组合算法[J].河北电力技术,2021,40(5):6-10.
WEN Yuxin, YANG Jun, ZHU Xu. Power grid unit commitment algorithm based on deep reinforcement learning[J]. Hebei Electric Power, 2021, 40(5): 6-10.

[90] JASMIN E A, AHAMED T P I, REMANI T. A function approximation approach to reinforcement learning for solving unit commitment problem with photo voltaic sources [C]// 2016 IEEE International Conference on Power Electronics, Drives and Energy Systems (PEDES), December 14-17, 2016, Trivandrum, India: 1-6.

[91] DE MARS P, O'SULLIVAN A. Applying reinforcement learning and tree search to the unit commitment problem [J]. Applied Energy, 2021, 302: 117519.

[92] 刘林鹏,朱建全,陈嘉俊,等.基于柔性策略-评价网络的微电网源储协同优化调度策略[J].电力自动化设备,2022,42(1):79-85.
LIU Linpeng, ZHU Jianquan, CHEN Jiajun, et al. Cooperative optimal scheduling strategy of source and storage in microgrid based on soft actor-critic [J]. Electric Power Automation Equipment, 2022, 42(1): 79-85.

[93] BENJAMIN K, BRI-MATHIAS S H, YINGCHEN Z. Autonomous energy grids [R]. Washington D C, USA: NREL Power Systems Engineering Center, 2017.

[94] RUELENS F, CLAESSENS B J, QUAIYUM S, et al. Reinforcement learning applied to an electric water heater: from theory to practice [J]. IEEE Transactions on Smart Grid, 2018, 9(4): 3792-3800.

[95] QIU S, LI Z, FAN D, et al. Chilled water temperature resetting using model-free reinforcement learning: engineering application[J]. Energy and Buildings, 2022, 255: 111694.

[96] 徐春雷,吴海伟,刁瑞盛,等.基于深度强化学习算法的"电网脑"及其示范工程应用[J].电力需求侧管理,2021,23(4):73-78.
XU Chunlei, WU Haiwei, DIAO Ruisheng, et al. Deep reinforcement learning-based grid mind and field demonstration application [J]. Power Demand Side Management, 2021, 23(4): 73-78.

[97] SONG Y H, HUANG S W, ZHANG Z M, et al. Risk assessment of power system cascading outages based on deep reinforcement learning [C]// 2021 40th Chinese Control Conference (CCC), July 26-28, 2021, Shanghai, China: 8273-8279.

[98] ZHANG Z M, YAO R, HUANG S W, et al. An online search method for representative risky fault chains based on reinforcement learning and knowledge transfer [J]. IEEE Transactions on Power Systems, 2020, 35(3): 1856-1867.

[99] BROCKMAN G, CHEUNG V, PETTERSSON L, et al. OpenAI Gym[EB/OL]. [2022-01-10]. https://arxiv.org/abs/1606.01540.

[100] HENRY R, ERNST D. Gym-ANM: open-source software to leverage reinforcement learning for power system management in research and education [J]. Software Impacts, 2021, 9: 100092.

[101] FAN T H, LEE X Y, WANG Y B. PowerGym: a reinforcement learning environment for volt-var control in power distribution systems[EB/OL]. [2022-01-10]. https://arxiv.org/abs/2109.03970.

[102] MAROT A, DONNOT B, DULAC-ARNOLD G, et al. Learning to run a power network challenge: a retrospective analysis [EB/OL]. [2022-01-10]. https://arxiv. org/abs/2103.03104v2.

[103] 陈海东,蒙飞,张越,等.基于生成对抗模仿学习的电力系统动态经济调度[J].电网技术,2022,46(11):4373-4380.
CHEN Haidong, MENG Fei, ZHANG Yue, et al. Dynamic economic dispatch of power system based on generative adversarial imitation learning[J]. Power System Technology, 2022, 46(11): 4373-4380.

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

[105] ZHANG H, CHEN H, BONING D, et al. Robust reinforcement learning on state observations with learned optimal adversary [EB/OL]. [2022-01-21]. https://arxiv.

org/abs/2101.08452.

冯　斌(1997—)，男，博士研究生，主要研究方向：人工智能在电力系统中的应用。E-mail：fengbinhz@zju.edu.cn

胡轶婕(2000—)，女，博士研究生，主要研究方向：人工智能在电力系统中的应用。E-mail：huyijie12210071@zju.edu.cn

黄　刚(1991—)，男，博士，副研究员，主要研究方向：机器学习、运筹优化等算法及其在网络化系统中的应用。E-mail：huanggang@zju.edu.cn

郭创新(1969—)，男，通信作者，博士，教授，主要研究方向：智能电网风险评估与调度决策、综合能源系统规划运行。E-mail：guochuangxin@zju.edu.cn

（编辑　顾晓荣）

Review on Optimization Methods for New Power System Dispatch Based on Deep Reinforcement Learning

FENG Bin[1], *HU Yijie*[1], *HUANG Gang*[2], *JIANG Wei*[1], *XU Huating*[1], *GUO Chuangxin*[1]

(1. College of Electrical Engineering, Zhejiang University, Hangzhou 310027, China;
2. Zhijiang Laboratory, Hangzhou 311121, China)

Abstract: With the continuous expansion of renewable energy integration scale, energy forms become more flexible and diverse, which presents new challenges to the dispatch operation of power systems. As the complexity and uncertainty of the system increase, the traditional optimization methods based on physical models are difficult to establish the accurate models for real-time and rapid solutions. In contrast, the deep reinforcement learning (DRL) can adaptively learn the scheduling strategies and make real-time decisions from historical experiences, avoiding the complex modeling process and coping with higher uncertainty and complexity in a data-driven manner. In this paper, the dispatch operation problems of new power systems are firstly introduced, then the principles and classification of DRL are described, and the advantages and disadvantages of various DRL methods to solve the dispatch decision problems of new power systems are analyzed. Finally, the trends that need further research are prospected.

This work is supported by National Natural Science Foundation of China (No. U22B2098), Zhejiang Provincial Natural Science Foundation of China (No. LQ20E070002), and National Scholarship Fund of China (No. 202106320157).

Key words: deep reinforcement learning (DRL); new power system; economic dispatch; optimal power flow; unit commitment