计算机应用研究
Application Research of Computers
ISSN 1001-3695,CN 51-1196/TP

《计算机应用研究》网络首发论文

题目：基于不确定性的深度强化学习探索方法综述
作者：逄金辉，冯子聪
DOI：10.19734/j.issn.1001-3695.2023.03.0130
收稿日期：2023-05-27
网络首发日期：2023-06-21
引用格式：逄金辉，冯子聪．基于不确定性的深度强化学习探索方法综述[J/OL]．计算机应用研究. https://doi.org/10.19734/j.issn.1001-3695.2023.03.0130

基于不确定性的深度强化学习探索方法综述

逄金辉，冯子聪

(北京理工大学 计算机学院，北京 100081)

摘 要：深度强化学习(Deep Reinforcement Learning，DRL)近年来在诸多复杂序列决策问题场景中，如游戏人工智能、无人驾驶、机器人和金融等领域中都取得了重要的成就。然而，在诸多现实场景中，深度强化学习的应用面临着采样成本高昂、效率低下的问题。场景中无处不在的不确定性是影响采样效率的重要原因，基于不确定性的深度强化学习探索方法成为解决上述问题的重要思想。该文首先简要介绍了深度强化学习中的重要概念和主流算法，列举了 3 种经典探索方法，并对这些方法面对复杂场景时的不足之处进行了总结；之后，介绍了不确定性的概念，以及将不确定性引入 DRL 探索问题研究的背景，在此基础上进行了归纳整理，将基于不确定性的探索方法分为了基于乐观性、基于环境不确定性、基于偶然不确定性 3 种形式，详细梳理了各类方法的基本原理和优缺点；最后，展望了基于不确定性的深度强化学习探索研究的挑战与可能的发展方向。

关键词：深度强化学习；探索；不确定性

中图分类号：TP391 doi: 10.19734/j.issn.1001-3695.2023.03.0130

Exploration approaches in deep reinforcement learning based on uncertainty: review

Pang Jinhui, Feng Zicong

(*School of Computer Science*, *Beijing Institution of Technology*, *Beijing* 100081,*China*)

Abstract: In recent years, deep reinforcement learning (DRL) has made significant achievements in many complex sequence decision problem scenarios, such as game artificial intelligence, unmanned driving, robotics and finance. However, in many real-world application, DRL faces the problem of high sampling cost and low sampling efficiency. The ubiquitous uncertainty in the scene is an important reason for affecting the problem, deep reinforcement learning exploration methods based on uncertainty have become an important idea to solve the above problems. First, it will briefly introduce the important concepts and mainstream algorithms of DRL. Then it list three classic exploration method, and discuss the shortcoming of these methods in complex scenarios. After that, this paper introduces the concept of uncertainty and the background of importing uncertainty into the research of DRL exploration problems. On this basis, it summarizes the existing exploration methods based on uncertainty, which are divided into three forms: optimism based, environmental uncertainty based and aleatoric uncertainty based approaches. It also analyze the basic principles, advantages and disadvantages of each methods in detail. Finally, this review prospects the challenges and possible development directions of DRL exploration based on uncertainty.

Key words: deep reinforcement learning (DRL); exploration; uncertainty

0 引言

近年来，深度强化学习[1]在围棋[2,3]、StarCraft II[4]、机器人[5]等领域取得了超越人类的表现。然而，DRL 中探索效率低下的问题[6]，使得即使在简单任务场景中，算法学习过程也需要巨大的样本量。例如，在简单的电子游戏 Atari 中，Rainbow DQN[7]采用多种提升训练速度的技巧，需要一千八百万帧才能达到人类玩家的水平，而普通人上手游戏只需要几分钟时间。现实世界中，场景的状态、动作空间都远大于 Atari 游戏，在具有稀疏奖励[8]、长交互步骤、噪声分布等特点的场景中，采样效率低下已经成为 DRL 研究中的瓶颈问题之一。

上述探索困难、采样效率低下的问题需要高效的 DRL 探索方法帮助解决，这对于提高算法效率，进一步扩展 DRL 的应用场景十分重要。研究者们为提升 DRL 探索能力做了大量工作，提出了基于内在动机[9]、模仿学习等探索方法，但这些方法通常泛化性能较差，或需要额外的人工奖励设计步骤。由于不确定性在深度学习(Deep Learning，DL)[10]领域的成功应用，这一方法也开始被应用于 DRL 探索方法的研究中，由于基于不确定性探索方法在不同环境之间优秀的泛化能力和可解释性，使其逐渐称为深度强化学习领域研究的热门方向。

1 深度强化学习基本理论

1.1 强化学习

强化学习(Reinforcement Learning，RL)[11]是不同于监督学习、无监督学习的第三种机器学习范式，通过智能体与环

收稿日期：2023-03-28；**修回日期：**2023-05-27

作者简介：逄金辉(1969-)，女，北京人，副教授，硕导，博士，主要研究方向为博弈论、多模态知识图谱；冯子聪(1993-)，男，山西永济人，硕士研究生，主要研究方向为深度强化学习、博弈论(397671030@qq.com).

境的交互进行学习，目标是使智能体在交互过程中获得最大的累计奖励[12]。强化学习问题一般被形式化为马尔可夫决策过程(Markov Decision Processes，MDPs)。一个 MDP 通常由状态空间 S，动作空间 A，状态转移矩阵 $P:S\times A\times S\to[0,1]$ 表示状态转移概率，奖励函数 $R:S\times A\times S\to\mathbb{R}$ 以及折扣因子 $\gamma\in[0,1]$ 共同组成，用五元组 $\langle S,A,P,R,\gamma\rangle$ 进行表示。强化学习的目标是最大化累积的奖励信号数值，即学习到一个策略 π，使得智能体在每一个时间步 t，通过观察当前环境所处的状态 $s_t\in S$，根据策略选择动作 $a_t\in A$，以一定的概率 P 到达下一状态 S_{t+1} 并获得对应奖励 $r_t\in R$。在这一智能体与 MDP 环境交互的过程中，根据连续行动的累积奖励调整智能体策略，实现累积的期望奖励值最大化的目标[13]。由于交互过程中累积奖励的不确定性，奖励期望使用折扣因子 γ 进行估计，学习的目标函数表示为

$$J(\pi)=\max_{\pi}\mathbb{E}_{a\sim\pi}\left[\sum_{i}^{T}\gamma^{i}R(s_i,a_i)\right] \tag{1}$$

1.2　深度强化学习

几乎所有的强化学习算法都依赖于价值函数的估计[12]。对于给定的策略 π，根据 1.1 节中的定义，智能体为每个状态定义一个价值函数 $V^{\pi}(s)=\mathbb{E}_{\pi}[G_t\,|\,S_t=s]$，类似地，为每个状态-动作对定义对应的价值函数 $Q^{\pi}(s,a)=\mathbb{E}_{\pi}[G_t\,|\,S_t=s,A_t=a]$，两个函数定义中 $G_t=\sum_{k=0}^{\infty}\gamma^k r_{k+t}$ 表示累积折扣奖励的期望值。经典的强化学习方法中常通过表格法(Tabular solution methods)或近似法(Approximate solution methods)对问题进行求解，在问题场景的动作状态空间较小时，将全部的状态值 $V^{\pi}(s)$ 或动作-状态值 $Q^{\pi}(s,a)$ 维护在一个表格中，根据实验结果对表格中对应价值进行更新，并通过贪婪原则确定最优策略下的动作。但在现实世界中，实际问题对应的状态与动作空间往往十分庞大，表格法难以对所有对应价值进行存储和更新，而传统的近似法也难以对非线性场景下的价值函数进行准确拟合。为了提升强化学习的策略学习效率和对实际问题的应用能力，研究人员将深度神经网络与强化学习相结合，借助深度学习中优秀的表征能力实现近似法中难以实现的各类函数近似，虽然结合深度神经网络的强化学习存在缺乏理论支撑和可解释性等问题，但利用深度学习的感知能力，DRL 解决了策略和值函数的建模与近似问题，提升了强化学习解决复杂问题、实现通用智能的能力，也成为了强化学习领域研究的主要趋势。

基于价值函数的定义，通过深度网络表示价值函数或策略的不同，大多数 DRL 算法能够分为两个不同的类型[14]：基于值的方法和基于策略的方法。

1.2.1　基于值的方法

基于值的方法通过深度神经网络学习近似的最优价值函数，隐式求解对应的最优策略。深度 Q 网络(Deep Q-Network，DQN)[15]是 Q 学习结合深度学习的代表性基于值的 DRL 算法，通过一个参数化的 Q 深度神经网络 $Q(s,a;\boldsymbol{\theta})$ 表示动作函数，取代传统 Q 学习算法中的 Q 值来完成学习。DQN 算法通过对采样得到的样本分批处理，最小化时间差分损失(TD-error)：$\mathcal{L}^{DQN}(\boldsymbol{\theta})=\mathbb{E}_{(s_t,a_t,r_t,s'_{t+1})\sim D}[y-Q(s_t,a_t;\boldsymbol{\theta})]^2$ 来完成参数 $\boldsymbol{\theta}$ 的学习，其中 D 表示样本缓存池，目标价值 $y=r_t+\gamma\max_{a'}Q(s_{t+1},a';\boldsymbol{\theta}^{-})$，其中 $\boldsymbol{\theta}^{-}$ 表示目标 Q 网络中固定的参数。从学习过程能看到 DQN 算法的两个主要特点：使用两个独立的 Q 网络，通过 Q 网络的参数 $\boldsymbol{\theta}$ 定期对目标 Q 网络中的参数 $\boldsymbol{\theta}^{-}$ 进行同步，使得学习过程中的目标函数保持稳定，增强了算法的鲁棒性；另一方面使用样本缓存池实现经验回放[16]尽可能消除样本的相关性，批训练过程中样本之间相互独立，进一步提升算法性能。基于 DQN 算法的经典值方法还有 Double DQN[17]，Dueling DQN[18]和 Rainbow DQN[7]等。

1.2.2　基于策略的方法

不同于基于值的方法，基于策略的方法直接对参数化的策略 π_{θ} 进行更新，以实现目标函数最大化的最优策略[19]。具体来讲，是将最大化期望收益的目标转换为基于策略的目标函数：$J(\theta)=\mathbb{E}_{\pi_{\theta}}[G_t]$，遵循策略梯度理论利用蒙特卡洛(Monte Carlo，MC)方法估计回报，实现参数化策略的更新迭代。代表性的策略梯度算法包括 REINFORCE 算法[20]，其学习过程中策略梯度方程可以表示为

$$\nabla_{\theta}J(\theta)=\mathbb{E}_{\pi_{\theta}}[\nabla_{\theta}\log\pi_{\theta}(a_t\,|\,s_t)Q^{\pi_{\theta}}(s_t,a_t)] \tag{2}$$

经典的策略算法还包括确定性策略梯度算法(Deterministic Policy Gradient，DPG)[21]。同时为解决策略梯度中高方差问题，一些算法中引入演员-评论家(Actor-Critic)结构，如优势演员-评论家算法(Advanced Actor-Critic，A2C)，异步优势演员-评论家算法(Asynchronous Advanced Actor-Critic，A3C)算法[22]等。在上述经典的基于策略的方法之上，近年来研究者们还提出了置信区间优化(Trust Region Policy Optimization，TRPO)[23]，近端策略优化(Proximal Policy Optimization，PPO)[24]，深度确定性策略梯度(Deep Deterministic Policy Gradient，DDPG)[25]等新算法，从不同方面提升基于策略方法的性能。

1.3　经典探索方法

智能体在于未知环境交互过程中，如何平衡探索和利用相互制约的困境，是强化学习范式最基本也是最重要的难题之一[12]。在传统强化学习研究中，通常通过随机扰动的贪婪策略实现一定程度的探索，此外，在多臂赌博机等简单场景中，研究者们还提出了许多具有理论保证的探索方法，按照方法不同的原理，大致可以分为以下几个主要类型：

1.3.1　贪婪探索

强化学习中最常见的探索策略就是贪婪探索：基于过去与环境交互过程建立的对环境模型的认知，贪婪地选择最大化期望收益的策略。贪婪探索还包括其他类型的改进方法，例如基于扰动的探索，包括著名的 ε 贪婪算法($\varepsilon-greedy$)，Boltzmann 策略；又例如确定等效控制等都属于贪婪探索方法[26]。

1.3.2　汤普森采样

汤普森采样(Thompson Sampling，TS)[27]是一个具有理论性能保证的探索策略，应用贝叶斯优化的思想，将目标函数方程看作一个随机分布：每次执行策略的过程中，首先从一个后验分布中对目标函数进行采样，在按照贪婪策略完成与环境的交互过程。该方法通过随机分布对未知环境的不确定性进行形式化，在量化不确定性的帮助下实现高效的环境探索。在采样过程之后，算法实现过程与贪婪探索一致，因此两类方法能够方便地实现替换或结合。

1.3.3　上置信界探索

上置信界(Upper Confidence Bound，UCB)探索[28]在探索过程中通过：

$$A_t=\arg\max_{a}\left[Q_t(a)+c\sqrt{\frac{\ln t}{N_t(a)}}\right] \tag{3}$$

进行动作选择，其中 $N_t(a)$ 表示动作在之前被选择的次数，超参数 c 控制探索程度。方法中的根号项形式化对价值函数估计的不确定性，构建一个对于动作价值潜在期望奖励的上置信界。在同等置信水平下，被访问次数更少的状态、

动作对应更大的置信区间，这表示智能体对环境更高的不确定性，通过对不同置信区间不同频率的探索，在降低对应不确定性的同时，也实现了相比于随机扰动更高效的探索行为。

1.3.4 经典探索方法的不足

随着深度强化学习成为强化学习范式领域研究的主流，经典探索方法面对更加复杂的环境，存在许多亟待解决的不足之处。虽然经典的强化学习探索方法，特别是 ε 贪婪探索方法现在依旧是 DRL 算法中应用最为广泛的探索方法，但显然对于基于随机扰动的贪婪探索而言，其并没有实现对环境中状态与动作的感知评估，这使得探索过程完全是随机的，这会导致在大规模空间、稀疏奖励等探索困难场景中探索低效甚至算法失效等问题。而基于 TS 或 UCB 的探索方法虽然在小规模场景中具有理论保证，但在 DRL 环境中，面临更加复杂、更具不确定性和动态性的场景往往难以直接应用，特别 UCB 方法的原理导致其很难在不同场景之间进行迁移[12]。考虑到 DRL 的发展与应用面临着非稳定状态、大规模状态与动作空间、稀疏奖励和噪声环境等新的挑战[29]，DRL 中的探索方法也亟需进一步深入研究。

2 深度强化学习中的不确定性

基于不确定性的方法在优化与决策问题中扮演着十分重要的角色。随着大数据时代人工智能领域的飞速发展，不确定性估计和度量方法在机器学习与深度学习领域得到了广泛的研究与应用，被用于处理数据与深度神经网络中的不确定性，以提升任务性能和模型的可解释性等[30]。而对于深度强化学习而言，不仅面临与深度学习相同的深度神经网络中的不确定性，还由于强化学习范式与未知环境交互的特性，引入了新的不确定性。因此对于 DRL 不确定性的研究与引入显得愈发重要。

深度强化学习中的不确定性一般被分为两个不同类型：偶然不确定性(aleatoric uncertainty)和认知不确定性(epistemic uncertainty)[31]。

2.1 偶然不确定性

偶然不确定性是由于环境和与环境交互过程中随机性带来的不确定性，虽然偶然不确定性能够通过方法进行评估建模，但不能被减少。由于偶然不确定性产生于强化学习智能体与环境的交互训练过程，因此其的重要性通常与具体应用场景十分相关，例如，围棋中不存在偶然不确定性，而扑克游戏中由于对手手牌未知，偶然不确定性是环境中的重要组成部分。强化学习过程中，偶然不确定性通常产生于三个主要环节(与 MDP 的重要组件相关)：随机性奖励信号，随机性观测以及随机性动作。奖励信号的随机性必然导致价值函数中出现不可降低的不确定性；观测随机性是由于环境部分可观测或状态转移方程中的随机性引起，在具有代表性的强化学习研究中环境通常都是不完美信息博弈场景，即部分可观测的情形，例如 StarCraft II，Dota 2 等环境。训练 StarCraft II 强化学习智能体的过程中不可能获取战争迷雾机制中的环境信息，这会导致不可削减的偶然不确定性。而在德州扑克，二十一点等游戏中，由于状态转移方程不是确定性方程而是随机概率方程，这导致在状态转移过程中会产生偶然不确定性；动作的随机性显然会导致后续状态的不确定性，例如在典型的随机策略算法(TRPO[23]，PPO[24]等)中，动作是从分布中进行采样，这会导致偶然不确定性的产生。尽管上述三种来源的偶然不确定性都能够一定程度被估计甚至进行量化，但偶然不确定性不同于认知不确定性，不可能对它进行削减。虽然如此，对偶然不确定性的感知依然在强化学习过程中有着重要的作用，高性能强化学习智能体的训练必然需要能够对高偶然不确定性和高认知不确定性的场景进行区分：例如，通常本文希望对不确定性高的环境区域进行更多的探索，但如果区域的不确定性是偶然不确定性，对区域更多的探索行为会导致整个训练过程的效率大大下降。因此虽然不可能被削减，但对于高性能强化学习算法而言依然需要关注对偶然不确定性的感知。

2.2 认知不确定性

在人工智能领域，特别是深度强化学习方向，不确定性与一些基础性的理论问题息息相关。而与偶然不确定性不同，模型训练过程中的认知不确定性是能够被削减的。比如简单的数字识别分类器，随着训练进行准确度提升，其认知不确定性是一个不断下降的过程。

在深度强化学习领域，例如在 DQN[15]算法中，强化学习智能体通过一个神经网络学习拟合价值函数，过程中一个常见的问题是 Q 值的过估计，这通常是由于估计过程中的样本存在噪声，且 $\mathbb{E}[\max Q] \geq \max \mathbb{E}[Q]$ 所导致的，最终将使得学习到的估计值高于真实值。这一问题表明降低认知不确定性并非能够通过简单的增大训练时间与空间开销来解决。实践过程中，DQN 算法在长时间训练后可能出现性能下降，即灾难性遗忘(catastrophic forgetting)[32]问题，即使在确定性环境中，认知不确定性的感知甚至量化依然是亟待研究的问题；

认知不确定性同样与强化学习中经典的探索-利用权衡问题存在深刻联系，智能体在于环境交互过程中对未知的环境和策略更多的探索行为可能会导致获得一个相较已知最优策略更低的奖励值，相对的如果利用已知策略获取当前最大奖励可能会导致错失潜在的更高奖励的策略。显然上述问题的解决，即高效的探索行动的实现过程中，认知不确定性是明确且重要的切入角度之一。对强化学习环境中，状态、动作或策略空间中的不确定性与对相应空间的探索紧密相关，高认知不确定性意味着智能体对于对应的策略或价值函数的不确定，而这通常是由于对相应空间探索行为的不充分导致的。这一关系也被显示或隐式地应用于深度强化学习的探索策略之中。

3 基于不确定性的深度强化学习探索方法

为解决面对大规模动作、状态空间以及稀疏奖励等场景下，深度强化学习领域采样效率低下的问题，研究者们提出了基于目标[33]、内在动机[9]和不确定性度量等深度强化学习探索方法。基于目标探索的方法通过对如何达到子目标的过程进行控制，以提高智能体在复杂环境中的探索效率。这一方法更多基于规划实现探索，关键在于对交互过程中的存储状态与轨迹信息的记录，并根据相关信息通过规划算法生成相应的子目标，最后通过学习如何完成子目标实现探索效率提升。基于内在动机的方法由生理学最初发展，并逐渐受到认知科学领域的关注，该方法从行为学和心理学中动机驱动高等生物自主探索未知环境的机理出发，将多种形式化的内在动机概念应用为智能体学习过程中的奖励信号，从而实现智能体在未知环境中的高效自主探索行为。基于内在动机的方法更多体现拟人化实现高效探索的思路。

与上述两类方法相比，本文关注的基于不确定性度量的方法主要基于不确定性优先(Optimism in the Face of Uncertainty，OFU)思想[34]。现实场景中各个领域都面临不确定性，包括金融投资，医疗诊断，体育比赛和气象预测等，

所有这些场景中，都包含基于收集到的数据，基于不确定领域的认知来实现预期目标的决策过程。基于人工智能系统，特别是强化学习进行决策被愈加广泛的应用于相关领域，为实现智能体的高效与可靠性，对不确定性的可信表示甚至量化显得越发重要。随着不确定性在监督学习领域的广泛研究与成功应用[30]，利用不确定性度量形成高效、高泛化性的探索方法也逐渐成为深度强化学习领域研究的热点。具体来说，基于不确定性度量的方法通常利用贝叶斯后验对认知不确定性形式化建模，或通过分布式价值函数将偶然不确定性加入学习训练过程，通过鼓励智能体更多地对高认知不确定性对应的动作与状态空间进行探索，同时尽可能降低在高偶然不确定性空间中的开销。

基于不确定性度量的方法中，智能体通常为状态，动作，价值函数，奖励或几项相结合的目标维护一个概率分布，并基于该分布选择相应的动作。根据不同的不确定性度量目标以及领域内的相关研究，一些研究中将不确定性度量按照在神经网络中实现方法分为 MC Dropout，基于自举的模型(Bootstrap Model)以及高斯混合模型(Gaussian mixture Model, GMM)[35]。本文主要聚焦于深度强化学习探索方向，基于探索过程中不确定性的不同角色，将基于不确定性度量的深度强化学习探索方法主要分为三类：基于乐观性的探索方法，基于环境不确定性的探索方法以及基于偶然不确定性的方法，如图 1 所示。

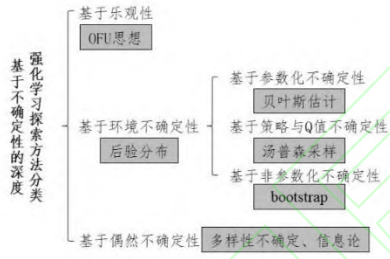

图 1　方法分类

Fig. 1　Method classification

3.1　常用测试环境

本节中首先整理介绍强化学习中常用的测试环境。本文主要介绍以下四种常见的开源环境。

3.1.1　Atari 游戏集

Atari 游戏集[36]包括 57 个不同的 Atari 街机游戏。这些游戏的状态空间通常是图片或者随机存取存储器(Random-Access Memory, RAM)快照。动作空间包含五类操作：上，下，左，右以及动作按键。Atari 游戏从探索难易程度能够大致分为两类：易于探索的 54 个游戏和探索困难的 3 个游戏[37]。在探索困难的游戏中，通常不存在容易获取的奖励信号，且在状态转移以及状态与奖励的相关性上也十分复杂。

3.1.2　VizDoom

VizDoom[38]是 Doom 游戏的一个 3 维仿真强化学习环境。作为第一人称视角游戏，玩家扮演角色所观察到的图像一般作为状态空间。动作空间包括一个八维度的方向控制以及两个动作按键：拾取动作与开门动作。该环境的主要优点是易用的场景编辑工具以及相对较低的计算开销。

3.1.3　Malmo

Malmo[39]是一个基于游戏我的世界(Minecraft)的强化学习环境。在游戏中，整个环境都是通过相同形状的砖块搭建而成。与前述 VizDoom 环境相同，Malmo 同样是第一人称视角，以角色观测图像作为状态空间。由于游戏环境搭建的特性，使得它具有在环境结构、空间尺寸、奖励函数等易于订制化的优点。

3.1.4　Mujoco

Mujoco[40]是一个基于物理仿真的常用环境。在强化学习场景中，Mujoco 通常被用作机器人相关的仿真工作。其常见的仿真对象包括猎豹，蚂蚁和类人生物等。强化学习的任务是在仿真场景中通过对对象关节等环节的控制实现行走、奔跑等一系列行为。

3.2　基于乐观性的探索方法

在基于乐观性的探索方法中，智能体遵循 OFU 思想，在探索过程中选择对应价值函数最高的动作进行行动，在价值函数估计的过程中，包含了方法对环境奖励的探索与预测，这一过程中的不确定性处理思路与方法，一般决定着方法在探索性能上的表现。

Jung 等人[41]提出利用高斯过程回归(Gaussian Processes Regression，GP)来提升强化学习探索过程采样效率的方法。通过一个基于 GP 的模型学习模块处理智能体与环境交互经验，对环境的状态转移与奖励方程进行灵活、准确的建模，再通过规划器完成最优策略选择。GP 的数学特性能够天然地呈现采样与预测之间的不确定性，在奖励方程较为简单的场景中，通过不确定性的引入以及模型学习与规划器的分离，提升了求解最优策略过程中的采样效率。Xie 等人[42]利用类似的方法，通过引入一个应用线性高斯模型表示的松弛项，来处理环境动力学建模过程中的不确定性，并在机器人控制的实际场景中验证了方法的有效性。该方法在能够更加灵活调整探索程度的同时，也能够在模型更新和学习过程中提高探索效率。虽然基于高斯过程的不确定性探索方法具有理论最优性保证等诸多优点，但基于模型的强化学习算法在大规模场景、可扩展性和计算效率等方面受限，且难以向非线性空间场景迁移。

D'Eramo 等人[43]将 Bootstrapped DQN 算法与汤普森采样进行结合，基于算法中的动作价值函数分布，使得汤普森采样的方法能够实现高维空间中的应用：在每一次动作选择之前，首先从分布中对每一个动作价值进行随机采样，之后执行采样结果中价值最高的对应动作，相比于 Bootstrapped DQN 文中提到的原方法及其改进的 Thompson DQN 方法，新的方法中不仅在每次动作执行前重新选择新的动作价值，且各个动作之间执行的是相互独立的采样过程。通过与汤普森采样方法更进一步的结合，为 Bootstrapped DQN 在更复杂环境下的性能表现实现了理论稳定性与实际性能的同步提升。

Osband 等人[44]基于最小二乘值迭代(Least-Squares Value Iteration，LSVI)方法的改进提出了随机化 LSVI(Randomized LSVI，RLSVI)方法，不同于低效的随机扰动探索，该方法主要基于汤普森采样的思想，RLSVI 通过对可能的价值函数首先进行采样，再根据采样得到的结果采取贪婪策略完成探索动作，这些分布式的价值函数可以看作对后验概率的一种近似。作者们还提供了方法理论保证的最差情形下近最优的遗憾值上界[45]，并在实验中证明该方法在线性 MDP 场景中大大提升了探索的效率。尽管受限于线性场景条件限制，使得 RLSVI 方法不能够应用于更大规模或连续空间场景，但为不确定性度量的探索方法提供了新的思路。

在 RLSVI 方法的基础上，Azizzadenesheli 等人[46]提出贝叶斯深度 Q 网络(Bayesian Deep Q-Network，BDQN)，通过在神经网络输出层结合一个贝叶斯线性回归模型，为函数近似的价值函数构建一个近似的后验估计，将汤普森采样的方法

与神经网络的非线性模型相结合，只在输出层的贝叶斯模型也不会影响神经网络本身的非线性表征能力，成功地将不确定度量的探索方法推广到非线性场景中。相比于 Double DQN[17]算法，BDQN 在诸多 Atari 游戏场景中实现了效率与性能的提升，部分游戏中得分提升了数倍甚至数十倍。但由于 BDQN 方法中在输出层前引入了新的模型参数，这可能会导致大规模任务中算法性能的不稳定性。

在上述基于 OFU 思想的方法中，智能体为应用该原则实现探索优化，都是通过对环境中的奖励进行一定程度的建模。一些方法中依托于基于模型的强化学习方法，还有一些并不直接对环境建模，而是通过依托于价值函数的分布估计实现对奖励的间接感知。显然基于模型的强化学习探索方法将受限于模型学习的精确性，特别是在探索困难的场景中，奖励的分布往往十分稀疏，这使得依托于价值函数的估计方法逐渐取代基于模型直接对奖励进行学习的方法，成为研究的主流。

3.3　基于环境不确定性的探索方法

在基于环境不确定性的方法中，智能体为环境的动作、状态等组成维护一个基于不确定性的分布，且不同于上述基于乐观性的探索方法，这类方法中智能体直接依据不确定性选择探索动作，以降低不确定性作为环境探索过程中的直接目标提升算法中的探索效率。这类方法又能根据具体原理与实现方法分为几个不同的类型。

3.3.1　基于参数化不确定性的探索方法

基于参数化不确定性的方法，为定义策略的相关参数引入不确定性，在策略学习的过程中，需要对不确定性的参数进行采样得到确定的策略后，才能实现后续整个学习的过程，在一定的学习周期之后，再根据学习的经验知识对参数以及其不确定性进行对应的更新，并不断重复这一过程直到达到性能需求。

Tang 等人[47]的工作非常直观的应用了这一思路，即在神经网络的参数上直接维护一个分布以体现参数的不确定性，将参数分布假定为一个高斯分布，通过采样实现探索中的动作选择，通过学习经验再实现分布的进一步估计。方法应用了参数化不确定性的思路，且由于是对网络参数本身的改进方法，能够比较容易地实现与不同深度强化学习算法的结合，但由于引入了较强的高斯分布假设，一定程度上限制了方法的性能表现和应用场景。

在 GEP-PG 方法[48]中，作者们受到演化理论的影响，基于策略梯度的算法，创新性地提出了一个两阶段组成的探索框架：首先通过一个基于好奇心的方法实现尽可能保证多样性的探索，之后将获取的经验存入回放池中，借助对专家知识的模仿学习来生成目标导向的最优策略。通过对探索过程和利用过程的解耦实现了更高的探索效率和更好的性能表现，并在稳定性上表现出优于传统策略梯度的优点。

Janz 等人[49]提出的 Successor Uncertainties 方法，结合可传递特征(successor features)与贝叶斯推理实现保证一致性的不确定性在整个样本中的传递性。可传递特征可以看作对状态的建模，通过贝叶斯线性模型在状态-动作对上实现不确定性度量，并最终一同构成价值函数的后验分布，同时应用时间差分方法更新的可传递特征，即后续可能状态-动作对不确定性的期望值，并通过该特征帮助构建原本独立的不确定性之间的相关性，最终探索过程中，通过从贝叶斯线性回归模型中进行参数的采样得到可传递特征的预测值。该方法不依赖于辅助状态或状态转移，具有易于实现和迁移的优点，但引入了新的可传递特征的不确定性。

在参数化后验方法的研究中，可以发现多数方法中都通过在线性 MDP 中引入基于贝叶斯回归的模型来体现与环境交互过程中的不确定性。在都依赖于贝叶斯模型的特点之外，3.2 节中的 RLSVI[41]、BDQN[46]等方法与上述 Successor Uncertainties[49]方法都是基于汤普森采样的思路，在构建的不确定性度量基础上实现 DRL 中探索效率的提升，这体现了汤普森采样的思想不同于 UCB 等传统探索方法，通过不确定性的度量能够自然地迁移到 DRL 算法中，使得传统方法能够在新的环境，特别是长交互步骤、高不确定性等环境中继续发挥重要作用。

3.3.2　基于策略与 Q 值的不确定性探索方法

不同于参数化不确定性方法中直接对不确定性建模的思路，基于策略与 Q 值的方法尝试将不确定性度量与强化学习范式中的价值函数相结合，智能体原本动作的 Q 值或固定策略被一个包含不确定性的分布替代，通过采样完成与环境交互的学习过程。

Stulp 的工作[50]是最早体现这一思路的研究之一，通过交叉熵方法(Cross-Entropy Method, CEM)控制算法学习过程中动作分布矩阵的协方差指标，以此实现不受动作空间维度限制的探索行为。该方法实际是在策略的参数空间上直接进行探索，并能随着场景规模的变化实现不同的探索率，在提升探索效率的同时，也保证了算法在不同环境中的健壮性和适用性。而在 Akiyama 等人[51]的工作中，基于重要性采样的思路，作者提出对探索过程的策略进行采样，在每一次与环境交互之前，实际策略是根据当前最优策略与过去采样经验通过优化算法计算得到的，这一定程度描述了策略的不确定性，并进一步提升了交互过程中获取的经验利用效率。在每次交互过程结束时，不仅对最优策略进行更新，同时需要计算、选择新交互过程中所需要的实际策略。这两类方法可以看作基于策略不确定性的早期工作。

在更一般的思路中，与直接参数化不确定性的探索方法相同，同样大量尝试利用贝叶斯框架来构建基于策略或 Q 值的不确定性度量。在 Strens 的工作[52]中，用一个贝叶斯后验分布取代通常的最大似然估计，并结合动态规划的方法实现分布的更新，在固定的一定时间内，通过采样得到的相应参数进行学习，并在时间结束时进行参数及其对应分布的更新，从而在探索的过程中保证一定的目标以及稳定性。但正如其他贝叶斯方法一样，特别是考虑到方法中贝叶斯框架与动态规划相结合的情况，该工作面临十分严重的计算开销问题。

同样基于贝叶斯框架，Guez 等人[53]的工作在前人研究的基础之上利用蒙特卡罗树搜索(Monte Carol Tree Search，MCTS)的方法，提升了贝叶斯框架下最优规划的计算效率。方法中将整个空间中的探索过程构建为一个树结构，同时为避免准确构建完整结构所需的巨大开销，通过采样与树搜索算法相结合的方式，构建近似的树结构，并使其相应状态分布尽可能地接近实际环境。通过采样与搜索算法的帮助，该方法大大提升了基于贝叶斯框架相关算法的可实施性，使其能够应用于更大规模、更复杂的问题场景之中。

Metelli 等人[54]的工作与 Successor Uncertainties 方法类似，同样关注到交互过程中不确定性的传递，并采用贝叶斯框架构建价值函数更新过程中不确定性的近似后验分布，但不同于上述的方法，在对后验分布更新时采用基于 Wassertein 重心的时间差分算法进行更新。基于上述思路，作者提出 Wasserstein Q 学习(Wasserstein Q-Learning，WQL)算法，提升了学习过程中样本的利用效率，并验证了算法在连续空间场

景下同样具有有效性。但由于引入了新的计算，WQL 方法在大规模场景中的计算开销是巨大的。

O'Donoghue 等人[55]提出不确定性贝尔曼方程(Uncertainty Bellman Equation，UBE)的概念，将量化的不确定性利用贝尔曼方程的形式联系起来，不再单单考虑某一个时间步上的不确定性，而是构建一个当前时间步未来继续执行时不确定性的期望，从而更有效地进行最优策略的探索。具体来说，方法中将 Q 值贝叶斯后验估计的方差作为不确定性的度量，并保持不确定性期望的计算与价值函数一致，这使得该方法的计算效率大大提高。但由于方法只计算贝叶斯估计的方差，因此不能采取类似汤普森采样的方法帮助探索，而是作为影响探索动作的附加项，以类似 UCB 方法的形式加入价值函数中。

然而，基于贝叶斯概率的方法都是基于状态或相应动作、策略进行后验概率分布的建模，在深度强化学习中，大规模的高维状态表征也需要神经网络来进行学习，这一过程中状态表示存在巨大的扰动，这也阻碍了类似方法在其基础之上构建后验概率分布的过程。如何解决类似方法与神经网络学习过程的冲突，是这一方向潜在的重要研究问题之一。

3.3.3　基于非参数化不确定性的探索方法

不同于参数化不确定性的思路，非参数化的方法中代表性的工作是基于自举(bootstrap)思路生成不依赖额外参数的后验估计，这类方法同样在线性 MDP 场景中取得了具有理论保证的探索性能提升。这一类中最具代表性的工作是 Bootstraped DQN[56]，基于深度强化学习经典的 DQN 算法，该方法中为每一个 Q 值维护若干个相互独立的估计器，在每次智能体与环境交互初始化时随机选择确定的估计器作为交互过程中的价值，这使得采样过程具有一定的一致性与稳定性，同时独立的估计器构成一个近似的价值函数后验概率分布，体现了应用汤普森采样处理不确定性以提升探索效率的思路。Bootstrapped DQN 与 RLSVI 在思路上十分接近，但通过自举的方法构建分布，这使得 Bootstrapped DQN 能够方便地结合到常见的深度强化学习算法中，特别是解决了 RLSVI 受限于线性 MDP 的问题，与非线性参数的价值函数相结合，在典型的难探索 Chain MDP 场景以及经典的 Atari 游戏场景中都实现了性能上的突破。Osband 等人[57]还在 Bootstrapped DQN 的基础上进行了改进，通过添加随机先验函数取代原本随机初始化的方法，能够增加不同估计器之间的差异，从而使得不确定性度量的估计更加准确，进一步提升了算法在不同不确定性场景之间的泛化性能。此外，还有许多利用类似思路结合自举方法与其他深度强化学习经典算法的研究，包括 Bootstrap 策略梯度[58]，Multi-DDPG[59]，MABDDPG[60]，以及 SOUP[61]，通过自举方法与策略梯度算法的结合，同样提升了学习过程中的探索效率。

Peer 等人[62]将 Bootstrapped DQN 与经典算法 Double DQN 的思路相结合，不同于应用独立的估计器来维护 Q 值的分布，该方法中利用相互独立的完整神经网络结构来实现分布的构建；与 Double DQN 方法相比，除了利用独立网络进行目标值预测外，在动作选择过程中，也利用了不同网络中的参数分布。实验表明，利用五个相互独立的网络，该方法能够实现类似 Rainbow DQN 的效果，而相比之下，该方法的思路与实现都更加简单直观。Hiraoka 等人[63]将 Bootstrapped DQN 与深度学习中常用的 MC Dropout 方法结合，同样是简单的思路与实现，该方法实现了改进后采样效率与计算开销上成倍的提升。

与上述提到的改进 Bootstrapped DQN 方法类似，Lee 等人[64]提出的 SUNRISE 方法同样是通过随机初始化的自举来度量学习过程中的不确定性，同时该方法中还应用加权贝尔曼备份的方法解决目标网络中误差传播过程中的不稳定性，最后基于 OFU 的思想，采用类似 UCB 算法应用网络方差决定探索动作的选择，通过对几类经典思路的有机整合，SUNRISE 方法能够通过不确定性估计在离散与连续场景中有效提升探索效率，并能够灵活、便利地与如 Rainbow DQN 等经典方法进行结合，实现性能的提升。

同样受到 Bootstrapped DQN 中多神经网络集成的思路启发，为解决每次采样单独模型可能导致的收敛性问题，Pearce 等人[65]的工作中采用多个模型共同拟合整体的状态分布情况。该方法中不同的模型被看作是一个高斯分布中对应的不同采样结果，采用贝叶斯推断保证模型的收敛性。不同于 Bootstrapped DQN，模型之间不存在相互独立的要求，而是整体满足一个高斯分布，这使得各个模型之间在保证整体收敛趋势的同时，保留了更多的多样性，同时更有利于通过不同初始化引入相应的先验知识。

Ciosek 等人[66]提出的 OAC 算法主要是针对 Actor-Critic 算法进行改进，面对低样本采样效率的问题，利用基于自举方法的网络在原有的近似低置信区间的基础上增加了一个高置信界，在探索过程中，通过置信区间上界确定探索的协方差已达到提升样本采样效率的目的，同时通过对探索策略与目标策略增加一个 KL 限制，实现探索效率提升的同时保证算法在学习过程中的稳定性。Bai 等人[67]提出的乐观性自举与反向推断(Optimistic Bootstrapping and Backward Induction，OB2I)方法通过逆向归纳的方法对整个交互经验中的不确定性进行度量，通过自举方法，为深度强化学习构建一个类似经典强化学习探索中 UCB 方法的框架。由于逆向归纳的引入，使得方法能够对整个交互周期中的不确定性进行一个保证时间一致性的度量，显然这能从理论的角度保障学习过程中探索效率的提升。OB2I 通过不确定性的逆向归纳使得自举方法与 UCB 能够有机地相结合，同时吸取了两者的理论上的优点，使得新的探索方法能够更多的考虑长交互环境中的不确定性，从而在探索效率上实现进一步的提升。

3.4　基于偶然不确定性的探索方法

对于存在大量随机性的环境，不确定性度量的过程中必须考虑两类不确定性的差别，即环境当前的不确定性可能不是由于探索的不充分，而是由于环境自身高随机性导致的偶然不确定性。由于偶然不确定性不可能在学习过程中通过算法设计等手段进行降低，因此若是不对不确定性的类型进行区分，对高偶然不确定性的空间依旧采取频繁的探索行为，会大大降低探索过程的效率与智能体的性能表现。为避免这一问题，许多工作在基于不确定性度量进行探索中同时考虑两类不确定性对学习过程的影响。

双层不确定性值网络(Double Uncertain Value Networks，DUVN)方法[68]通过两个网络实现两类不确定性的独立度量，一方面通过贝叶斯 Dropout 来度量认知不确定性，另一方面同样借助贝尔曼方程的计算在返回值上构建一个高斯概率分布，以此衡量场景中的偶然不确定性。结合分别度量的两类不确定性，构成一个价值网络，并根据该网络分布，通过托马斯采样的方法生成最优策略。然而，虽然 DUVN 方法中实现了基于偶然不确定性的探索，对两类不确定性分别进行了度量，也在实验中实现了一些场景中探索效率的提升，但它并没有针对偶然不确定性引发的问题做合适的处理，即不能

解决高偶然不确定性环境中乐观探索带来的效率与性能损失的问题。

Nikolov 等人[69]的工作受启发于信息导向采样(Information Directed Sampling，IDS)方法[70]在单臂老虎机环境中的性能表现，基于该思想进一步将 IDS 方法扩展到了更一般的 MDP 场景中，应用 IDS 方法能够同时实现探索过程中对认知不确定性与偶然不确定性的度量，具体来讲，方法中通过 Bootstrapped DQN 帮助度量认知不确定性，同时通过分布式强化学习中的 C51 方法[71]来度量偶然不确定性的近似分布。在进行探索的过程中，方法分别根据两类不同的不确定性构建对应的遗憾值和信息增益值，最终通过遗憾值与信息增益的比值决定探索方向。通过对两类不确定性进行不同处理再有机地结合，是基于 IDS 的强化学习探索方法最具创新性的突破。在 Clements 等人[72]的工作中，同样对两类不确定性分别进行了处理。该方法基于贝叶斯框架，通过对学习过程中返回值的分布进行学习，并将其方差分离出来作为偶然不确定性的度量指标，相应的均值被看作认知不确定性的度量，方法中采用两个平行的神经网络来实现不确定性的学习，不同于 IDS 中采用不同方法度量两类不确定性，该方法中通过两个同质化的分位数回归深度 Q 网络(Quantile Regression Deep Q-Network，QR-DQN)[73]实现两类不确定性的度量，其度量结果也能够应用与 IDS 方法并在实验中取得了类似的性能。Pearce 等人[74]的工作中，通过最大后验估计(Maximum Approximate Posteriori，MAP)采样来为神经网络生成两组不同的参数。借助参数之间的均方误差来度量估计认知不确定性，学习过程中随着神经网络的收敛，参数估计的方差下降，这也代表对应的认知不确定性有效地降低。同时两组参数估计之间的协方差用来表示偶然不确定性。同时，参数估计只对动作选择过程产生影响，即不确定性没有参与神经网络中其他训练过程。实验证明，该方法能够有效地提升智能体的性能，由于良好的解耦设计，使得方法能够更有效的适应各种不确定性场景。这些分别处理两类不确定性的方法对实现实际场景下各类复杂领域的应用，例如意外监测等有着重要意义，但显然这些方法都需要额外的计算开销，甚至网络结构的增加作为代价。

在衰退方差左截断(Decaying Left Truncated Variance，DLTV)方法[75]中，同样考虑到两类不确定性对于探索过程的不同影响，一方面通过延时处理偶然不确定性的影响，另一方面通过学习分布作为奖励项利用认知不确定性提升探索效率。同样以分布式强化学习方法 QR-DQN 为基础，DLTV 相比起上述其他同时考虑两类不确定性的方法，它并没有显示计算两类不确定性的过程，而是直接采用学习到的相应分布，将其分布方差看作天然的不确定性量化指标，同时结合延时奖励的方式，在应用认知不确定性帮助探索的过程中，尽可能地避免了环境中偶然不确定性带来的影响。显然 DLTV 方法在计算效率上由于前述的其他需要额外计算的方法，实验中，DLTV 不仅实现了采样效率的提升，还进一步提升了探索过程的安全性。在 DLTV 方法的基础上，针对分布式强化学习中回归过程数值非单调的问题，从理论角度出发，Zhou 等人[76]提出无交错的分位数回归(Non-Crossing Quantile Regression, NC-QR)方法，在对估计过程增加单调性限制后，避免上述非单调问题的出现。NC-QR 方法从理论层面保证了分布估计过程的准确性与可解释性，保证方法在无限次迭代中可以收敛至固定点，在 DLTV 的基础之上进一步提升了探索过程的效率与稳定性。

Mai 等人[77]的工作从监督学习(supervised Learning，SL)的相关问题与算法中得到启发，并将其推广到 DRL 的探索问题中。算法中利用一个称作倒方差(inverse variance)的方法实现两类不确定性的解耦估计。详细来讲，对每次批处理过程中的值函数损失，将其分解为两项：批处理倒方差(Batch Inverse Variance，BIV)损失 $\mathcal{L}_{BIV}$，其本质是标签中的噪声方差，即环境中的偶然不确定性，对噪声越大的交互数据，利用更小的权值抵消噪声对学习的影响；以及方差网络(variance networks)损失：$\mathcal{L}_{LA}$，通过神经网络估计噪声的规模，利用最大似然估计(Maximum Likelihood Estimation，MLE)损失得到对认知不确定性的估计。通过集成的方差网络，结合自举方法能够实现对两类不确定性的估计与应用。这一方法不仅能与 DQN 算法，也能够与 SAC 等方法相结合，都在采样效率上得到了很大的提升。

与上述方法不同，Mavor-Parker 等人[78]的工作中尝试仅通过偶然不确定性的估计实现探索性能的提升，通过一个深度神经网络预测状态表示，并用估计值方差表征偶然不确定性，并通过 MLE 损失对不确定性进行学习与更新。文献[79]实现了几乎相同的方法，并尝试从理论角度对两类不确定性的解耦以及不确定性估计与探索的关系进行了讨论。

尽管相关研究在利用偶然不确定性进行探索的方向取得了一些成果，偶然不确定性的估计始终缺少理论支持，这也使得相关方法的可靠性与鲁棒性缺乏保障。

3.5　总结

表 1 中本文对主要的基于不确定性探索方法进行了总结对比。不确定性可以纳入大多数深度强化学习探索方法之中，而不是完全独立于传统方法的独立评估，这使得可以通过一些简单的、在监督学习领域已有大量工作基础之上的技术来实现较高的收益。同时，不同类型的不确定性及其度量方法，能够帮助在具有不同不确定性水平的不同环境中，基于不同类型的基础算法实现跨类别的探索方法改进。

4　基于不确定性探索问题的关键问题及挑战

尽管基于不确定性的探索方法研究在 DRL 领域已经取得了一些成果，一定程度上提升了智能体对未知环境的探索效率与学习速度，但是也面临许多亟需解决的问题和挑战。通过分析对于未来研究面对的关键挑战的特性与成因，尝试找出解决问题的核心因素。

a)大规模状态-动作空间中的探索[29]。在面临状态、动作空间越来越大的问题是，探索方法通常也需要借助规模不断增大的神经网络来完成不确定性的度量工作。在本文介绍的 Bootstrapped DQN[56]，OAC[66]和 IDS[69,70]等方法中，无论通过自举方法或是贝叶斯网络来实现概率分布的拟合，都需要消耗大量的计算资源，在实际工作中通常也需要结合一些优化方法来保证计算的效率。从理论角度来看，通过自举方法实现真正的价值分布拟合需要无穷多次单个价值函数的计算，这显然是不可能实现的，实际实现过程中一般通过十个等概率的价值函数拟合实际的分布情况，显然更多的计算能够获取更加准确的分布，但这同时受到计算资源的限制。如何通过一定的方法解决不确定性度量方法中精确程度与计算开销的均衡，是一个潜在的研究方向。一个可能的解决方法是，通过引入表示学习中的相关方法，在不确定性度量的过程中借助表示学习构建更有效的状态表征，针对更加关键的状态进行不确定性的度量与探索工作，不在相关性较低的状态与动作空间中消耗太多的计算资源，通过更加显式、更具可解

释性的表示学习方法，从而提升智能体在大规模场景中探索、学习的效率。

除了大规模的状态空间所带来的问题，大规模的动作空间也是阻止大多数深度强化学习方法在实际场景中落地的关键挑战。不同于常见实验环境中离散、小规模的动作空间设置，实际问题中的动作空间经常包含大量的离散动作，或是包含复杂结构的复合动作的大规模离散，甚至连续空间。这常常导致在实验中性能良好的传统算法在实际场景中性能的大幅下降甚至完全失效，因此，在探索方法的研究过程中也应该更多的考虑大规模动作空间下的相关问题。可惜的是，目前在这一方向并没有太多的相关研究。一些研究提出通过分层强化学习(hierarchical reinforcement learning, HRL)[80]的方法将大规模、复杂的动作空间进行一定的抽象与分层来解决探索困难的问题；另一些研究提出通过动作语义来帮助动作空间结构和表征的学习，通过近似的语义帮助提升智能体的知识学习能力与泛化能力，同时通过语义相差较大的动作帮助实现大规模动作空间中的高效探索。虽然这些方法从不同的角度尝试缓解大规模动作空间场景中所面临的问题，但总体来说，该场景下的探索方法依然无法满足任务的需求，是探索领域未来重要的研究问题。

b)延迟与稀疏奖励场景下的探索[29]。尽管基于不确定性度量以及其他原理的探索方法在稀疏、延迟奖励场景中取得了一些研究成果[55~57]，但当前的方法面临这类场景问题，特别是面对长交互步骤的情形时，依然存在许多不足。特别是对基于不确定性度量的探索方法而言，虽然通常在任务之间表现出更优秀的泛化性，但在面对这类稀疏、延迟奖励的场景时，性能通常相比基于内在激励等其他原理的探索方法有所不如。尽管一些工作在类似 Motezumas’ Revenge 这样典型场景中取得了超越人类玩家的性能结果，但这些工作的探索效率不能让人满意，而且由于引入了大量环境相关的约束条件，也使得这些方法严重缺乏泛化能力。面对向实际场景应用的需求，面对长交互步骤的实际场景，往往需要将探索方法与环境的先验知识或内在动机相结合，显然这些信息的获取需要付出高昂的成本，并且不论以何种形式对先验知识进行表示，其与现有的探索方法相结合也需要进一步研究。总体来说，面向延迟、稀疏奖励，特别是长交互步骤场景下的探索依然是待解决的研究问题，同时也是实现深度强化学习实际应用过程中的重要环节。

c)噪声干扰场景下的探索[6]。现实世界通常伴随高随机性，这使得智能体与环境交互过程中观测结果与动作空间都存在不可预测的情况。3.4 节中基于偶然不确定性的探索方法是尝试通过偶然不确定性的形式化来解决这一问题。此外，也有一些研究尝试通过注意力机制等方法构建一个更紧凑的特征表示，通过舍弃弱相关高随机性的特征，构建噪声不敏感的智能体面对这类场景。但基于不确定性的探索方法大多针对环境状态的随机性，对动作空间中的噪声并没有得到充分研究。此外，当前研究一般通过在 Atari 环境中添加随机噪声等方法模拟实际噪声环境，但对实际场景问题中噪声的感知与建模显然复杂度更高，这也是探索方法未来研究的方向之一。随着大规模生成模型、扩散模型等研究日益成熟，通过对抗训练等方法的帮助来实现更丰富、更真实的噪声场景，是得到更具鲁棒性的探索方法的可能思路。

d)泛化性能[6]。大多数探索方法通常只能够在训练的场景下完成相应的任务，而不具备在不同场景、不同域之间的迁移与泛化能力。这方面，基于不确定性度量的探索方法展现出优于其他类型探索方法的潜力，通常能够在不同的任务之间展现一定的通用性和泛化性。在面对新环境时，泛化性优异的方法能够更好的利用已学习的经验与知识实现新的任务，这将大大降低智能体学习训练过程的计算开销。同时，具有高泛化性能的智能体在面临类似场景，例如相同场景下不同搜索目标，或相同任务目标更大状态空间等情形时，也能够更好的发挥作用。

表 1　主要不确定性探索方法对比

Tab. 1　Comparison of mainly uncertainty approaches

探索方法	基础算法	方法分类	方法性能*(实验环境：场景 得分(基线方法 得分))*	动作/状态空间
Jung 等[41]	-	基于乐观性	Mujoco：Inverted Pendulum 0(SARSA -10)	离散/连续
Xie 等[42]	MPC	基于乐观性	机器人手部机械仿真：完成十个预定动作	连续/连续
D’Eramo 等[43]	Bootstrapped DQN	基于乐观性	Mujoco：acrobot -100(Thompson DQN -120)	连续/连续
Osband 等[44]	LSVI	基于乐观性	Tetris：5000(LSVI 4000)	离散/离散
Bootstrapped DQN[56]	DQN	基于环境不确定性	Atari：James Bond 1000(DQN 600)	离散/离散
Tang 等[47]	DDPG	基于环境不确定性	Mujoco：sparse mountain car 0.2(NoisyNet 0)	离散/连续
GEP-PG[48]	DDPG	基于环境不确定性	Mujoco：Half Cheetah 6000(DDP 5445)	连续/连续
Successor Uncertainties[49]	DQN	基于环境不确定性	49 Atari games：77.55%超越人类表现(Bootstrapped DQN 67.35%)	离散/离散
Stulp 等[50]	PI^2	基于环境不确定性	Ball batting：20 步学习过程	连续/连续
Akiyama 等[51]	LSPI	基于环境不确定性	Ball batting 2 DoF simulation：67(Passive learning 61)	离散/连续
Strens 等[52]	DP	基于环境不确定性	Maze：1864(QL SEMI-UNIFORM 1147)	离散/离散
Guez 等[53]	Policy learning	基于环境不确定性	Dearden Maze：965.2(SBOSS 671.3)	离散/离散
UBE[55]	DQN	基于环境不确定性	Atari：Montezuma Revenge 3000(DQN 0)	离散/离散
Nikolov 等[69]	Bootstrapped DQN C51	基于偶然不确定性	55 atari games：1058%得分基于人类表现(C51 701%)	离散/离散
QR-DQN[73]	DQN	基于偶然不确定性	57 atari games：915%得分基于人类表现(DQN 228%)	离散/离散

5　基于不确定性探索方法研究展望

在不确定性强化学习探索方法的研究中，其高效的探索性能和优秀的泛化性能吸引了许多研究人员的注意，基于优化与深度学习领域不确定性度量的研究，帮助智能体借助环境、动作空间或深度网络中的不确定性，有效提升 DRL 方法的探索能力。当前基于不确定性的探索方法仍有许多发展潜力，为更好适应 DRL 面向实际场景的应用需求，不仅需要解决已有

的挑战和问题，还为方法进一步研究方向提出以下展望：

a)收敛性[81]证明。在基于不确定性度量的探索方法中，不论基于乐观估计或是汤普森采样的不同方法，理论上为实现最终向最优策略收敛的结果，都需要算法能够保证不确定性在学习过程中向零收敛。一些相关工作已经能够证明，在简单的线性 MDP 场景下，认知不确定性能够确保收敛到零这一条件。在非线性约束的 MDP 场景中，理论上同样会随着智能体对环境的不断充分探索，相应的鼓励探索的不确定性应当都会趋近于零，同时构建的后验分布会进一步逼近真实的分布。但实践中由于维度爆炸问题，深度强化学习中通常利用神经网络拟合函数代替原本的不确定性估计，这可能会导致不确定性度量过程中很难实现最终的收敛。不仅如此，实际许多工作中并不计算实际的分布，而是通过一定方法在计算开销允许的情况下尽可能拟合准确的分布，这同样使得对不确定性度量的置信度与收敛性几乎不能够被准确的衡量。在未来研究中，如何更准确同时更高效地完成对不确定性度量和后验概率分布构建，是需要关注的方向，在研究中不仅需要关注实际算法的性能表现，也要保证方法的理论完备性。

b)安全探索[82]。对于实际场景的应用，安全探索是十分重要的要求。然而在目前的研究中，有关于安全探索的研究并不多见，实践过程中往往通过仿真来避免实际场景下的开销，或者需要严重依赖人工订制的规则来避免实际场景中的各类探索安全问题[83]。显然，当前的不确定性深度强化学习探索方法，即时通过各类约束条件的限制依然很难避免实际应用中各类以外情况的发生，这严重阻碍了深度强化学习实际场景中的应用。为解决这一问题，智能体应当能够识别并处理各类不安全场景，而不是依赖于人工订制的规则。实际情况中，人工制定的规则也很难能够覆盖所有可能的意外情形，如何通过安全探索方法的研究，帮助智能体通过强化学习的方法安全地获取处理各类意外情景的能力，是未来研究的方向之一。

c)多智能体强化学习[84]的探索。虽然多智能体强化学习在游戏智能体等领域取得了一定的成果，但该场景下的探索方法研究仍处于起步阶段。对多智能体强化学习场景中的探索而言，随着智能体数目的增加，其相应的状态动作空间也迅速扩张，这对于基于不确定性度量的方法带来了计算开销和收敛性等问题，此外，由于多智能体环境新的特点，也为探索方法的研究带来了新的问题：首先是多智能体场景下非稳定状态、部分可观测环境的特点，这会导致现有基于单智能体的探索方法性能下降甚至完全失效；其次是多智能体之间的合作与对抗问题，特别在合作场景中，多智能体往往共享奖励信号，这时需要合理的奖励信号分配机制或其他方法实现每个智能体探索动作的选择，这超出了独立处理智能体的传统方法的能力，而受限于部分可观测环境限制，多智能体的合作往往难以达成，这进一步限制了智能体对环境的探索效率；最后，如何在分布式执行的框架下实现智能体的高效探索，由于在局部观测条件下，智能体的状态不仅与环境交互，实际也受到了其他智能体动作的影响，对于实际奖励信号和由其他智能体动作引起的伪奖励的区分，会严重影响探索方法的性能效率。

d)离线强化学习中的不确定性利用。近年来离线强化学习(Offline Reinforcement Learning)[85]逐渐受到强化学习研究人员的重视，其数据驱动的特性，使得该范式面临分布外泛化(Out-of-Distribution，OOD)问题[86]的困扰。面对数据的不确定性，许多研究工作尝试从基于不确定性的探索方法出发，解决 OOD 问题。例如 An 等人[87]的工作利用自举方法估计不确定性，基于松弛演员评论家(Soft Actor-Critic，SAC)算法[88]，通过增加 Q 值函数的方式实现不确定性的引入，最终提升算法在 OOD 条件下的性能。还进一步提出集成多样化演员评论家(Ensemble-Diversified Actor Critic，EDAC)算法[89]，通过不确定性的引入提升 Q 值在学习过程中的方差，以此应对 OOD 数据带来的探索与学习困难问题。令人惊喜的是，虽然只是简单地添加不确定性估计，但该方法能够有效提升离线强化学习的算法性能。Lee 等人[90]为了使离线学习策略能够更好地迁移到在线学习过程中，同样利用不确定性帮助，避免了在线学习初期的自举过程破坏离线学习中得到的优秀初始策略，使得算法在迁移过程中更具鲁棒性。由于不需要额外的探索，因此离线强化学习方法能够显著地降低学习过程中的风险与学习成本，同时数据驱动的模式能够帮助强化学习范式更好的利用人工智能领域丰富的数据资源，这也是离线强化学习最主要的优势。随着离线强化学习的进一步研究，基于不确定性的方法也显露出在 OOD 场景下提升算法性能的巨大潜力。

6　结束语

本文主要对基于不确定性度量的深度强化学习探索方法研究进展做了综述。首先简单介绍了强化学习的基本概念、经典算法、常见的探索方法及其优缺点，紧接着介绍了不确定性的概念及其在深度强化学习领域引入的背景，并基于此对基于不确定性度量的探索方法做了详细的梳理与介绍。主要包括这样几个类型：基于乐观性的探索方法，基于环境不确定性的方法以及考虑偶然不确定性的探索方法。最后，对该领域内亟待解决的相关问题和未来可能的研究方向进行了整理。

参考文献：

[1] 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.

[2] 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.

[3] Silver D, Schrittwieser J, Simonyan K, *et al.* Mastering the game of go without human knowledge [J]. nature, 2017, 550 (7676): 354-359.

[4] Vinyals O, Babuschkin I, Czarnecki W M, *et al.* Grandmaster level in StarCraft II using multi-agent reinforcement learning [J]. Nature, 2019, 575 (7782): 350-354.

[5] Andrychowicz O A I M, Baker B, Chociej M, *et al.* Learning dexterous in-hand manipulation [J]. The International Journal of Robotics Research, 2020, 39 (1): 3-20.

[6] Ladosz P, Weng L, Kim M, *et al.* Exploration in deep reinforcement learning: A survey [J]. Information Fusion, 2022.

[7] Hessel M, Modayil J, Van Hasselt H, *et al.* Rainbow: Combining improvements in deep reinforcement learning [C]// Proceedings of the AAAI conference on artificial intelligence. 2018, 32 (1) .

[8] 杨惟轶, 白辰甲, 蔡超, 等. 深度强化学习中稀疏奖励问题研究综述 [J]. 计算机科学, 2020, 47 (3): 182-191. (Yang Weiyi, Bai Chenjia, Cai Chao, *et al.* Survey on sparse reward in deep reinforcement learning [J]. Computer Science, 2020, 47 (3): 182-191)

[9] 曾俊杰, 秦龙, 徐浩添, 等. 基于内在动机的深度强化学习探索方法

综述 [J/OL]. 计算机研究与发展: 1-24 (2022-09-17) [2023-03-22] https://kns. cnki. net/kcms/detail/11. 1777. TP. 20220916. 1221. 002. html (Zeng Junjie, Qin Long, Xu Haotian, *et al.* Exploration approaches in deep reinforcement learning based on intrinsic motivation: a review [J/OL]. Computer Research and Development: 1-24 (2022-09-17) [2023-03-22] https://kns. cnki. net/kcms/detail/11. 1777. TP. 20220916. 1221. 002. html.)

[10] LeCun Y, Bengio Y, Hinton G. Deep learning [J]. nature, 2015, 521 (7553): 436-444.

[11] 闫友彪, 陈元琰. 机器学习的主要策略综述 [J]. 计算机应用研究, 2004 (7): 4-10. (Yan Youbiao, Chen Yuanyan. A survey on machine learning and its main strategy [J]. Application Research of Computers, 2004, 21 (7): 4-10)

[12] Sutton R S, Barto A G. Reinforcement learning: An introduction [M]. MIT press, 2018.

[13] 高阳, 陈世福, 陆鑫. 强化学习研究综述 [J]. 自动化学报, 2004 (01): 86-100. (Gao Yang, Chen Shifu, Lu Xin. Research on reinforcement learning technology: a review [J]. Acta Automatica Sinica, 2004 (01): 86-100)

[14] 刘全, 翟建伟, 章宗长, 等. 深度强化学习综述 [J]. 计算机学报, 2018, 41 (1): 1-27. (Liu Quan, Zhai Jianwei, Zhang Zongchang, *et al.* A survey on deep reinforcement learning [J]. Chinese Journal of Computers, 2018, 41 (1): 1-27)

[15] Mnih V, Kavukcuoglu K, Silver D, *et al.* Playing atari with deep reinforcement learning [EB/OL]. (2013-12-19) [2023-03-08]. https://arxiv. org/pdf/1312. 5602. pdf.

[16] Lin L J. Reinforcement learning for robots using neural networks [M]. Carnegie Mellon University, 1992.

[17] Van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double q-learning [C]// Proceedings of the AAAI conference on artificial intelligence. 2016, 30 (1) .

[18] Wang Z, Schaul T, Hessel M, *et al.* Dueling network architectures for deep reinforcement learning [C]// International conference on machine learning. PMLR, 2016: 1995-2003.

[19] Sutton R S, McAllester D, Singh S, *et al.* Policy gradient methods for reinforcement learning with function approximation [J]. Advances in neural information processing systems, 1999, 12.

[20] Williams R J. Simple statistical gradient-following algorithms for connectionist reinforcement learning [J]. Reinforcement learning, 1992: 5-32.

[21] Silver D, Lever G, Heess N, *et al.* Deterministic policy gradient algorithms [C]// International conference on machine learning. Pmlr, 2014: 387-395.

[22] Mnih V, Badia A P, Mirza M, *et al.* Asynchronous methods for deep reinforcement learning [C]// International conference on machine learning. PMLR, 2016: 1928-1937.

[23] Schulman J, Levine S, Abbeel P, *et al.* Trust region policy optimization [C]// International conference on machine learning. PMLR, 2015: 1889-1897.

[24] Schulman J, Wolski F, Dhariwal P, *et al.* Proximal policy optimization algorithms [EB/OL]. (2017-08-28) [2023-03-08]. https://arxiv. org/pdf/1707. 06347. pdf.

[25] Lillicrap T P, Hunt J J, Pritzel A, *et al.* Continuous control with deep reinforcement learning [EB/OL]. (2019-07-05) [2023-03-08]. https://arxiv. org/pdf/1509. 02971. pdf.

[26] Curi S. Epistemic Uncertainty for Practical Deep Model-Based Reinforcement Learning [D]. ETH Zurich, 2022.

[27] Thompson W R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples [J]. Biometrika, 1933, 25 (3-4): 285-294.

[28] Auer P, Cesa-Bianchi N, Fischer P. Finite-time analysis of the multiarmed bandit problem [J]. Machine learning, 2002, 47: 235-256.

[29] Yang T, Tang H, Bai C, *et al.* Exploration in deep reinforcement learning: a comprehensive survey [EB/OL]. (2023-02-02) [2023-03-37]. https://arxiv. org/pdf/2109. 06668. pdf.

[30] Abdar M, Pourpanah F, Hussain S, *et al.* A review of uncertainty quantification in deep learning: Techniques, applications and challenges [J]. Information Fusion, 2021, 76: 243-297.

[31] Hüllermeier E, Waegeman W. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods [J]. Machine Learning, 2021, 110: 457-506.

[32] French R M. Catastrophic forgetting in connectionist networks [J]. Trends in cognitive sciences, 1999, 3 (4): 128-135.

[33] Guo Y, Choi J, Moczulski M, *et al.* Memory based trajectory-conditioned policies for learning from sparse rewards [J]. Advances in Neural Information Processing Systems, 2020, 33: 4333-4345.

[34] Lai T L, Robbins H. Asymptotically efficient adaptive allocation rules [J]. Advances in applied mathematics, 1985, 6 (1): 4-22.

[35] Hubschneider C, Hutmacher R, Zöllner J M. Calibrating uncertainty models for steering angle estimation [C]// 2019 IEEE intelligent transportation systems conference (ITSC) . IEEE, 2019: 1511-1518.

[36] Bellemare M G, Naddaf Y, Veness J, *et al.* The arcade learning environment: An evaluation platform for general agents [J]. Journal of Artificial Intelligence Research, 2013, 47: 253-279.

[37] Aytar Y, Pfaff T, Budden D, *et al.* Playing hard exploration games by watching youtube [J]. Advances in neural information processing systems, 2018, 31.

[38] Kempka M, Wydmuch M, Runc G, *et al.* Vizdoom: A doom-based ai research platform for visual reinforcement learning [C]// 2016 IEEE conference on computational intelligence and games (CIG) . IEEE, 2016: 1-8.

[39] Johnson M, Hofmann K, Hutton T, *et al.* The Malmo Platform for Artificial Intelligence Experimentation [C]// IJCAI. 2016: 4246-4247.

[40] Todorov E, Erez T, Tassa Y. Mujoco: A physics engine for model-based control [C]// 2012 IEEE/RSJ international conference on intelligent robots and systems. IEEE, 2012: 5026-5033.

[41] Jung T, Stone P. Gaussian processes for sample efficient reinforcement learning with RMAX- like exploration [C]// Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part I 21. Springer Berlin Heidelberg, 2010: 601-616.

[42] Xie C, Patil S, Moldovan T, *et al.* Model-based reinforcement learning with parametrized physical models and optimism-driven exploration [C]// 2016 IEEE international conference on robotics and automation (ICRA) . IEEE, 2016: 504-511.

[43] D'Eramo C, Cini A, Restelli M. Exploiting action-value uncertainty to drive exploration in reinforcement learning [C]// 2019 International Joint Conference on Neural Networks (IJCNN) . IEEE, 2019: 1-8.

[44] Osband I, Van Roy B, Wen Z. Generalization and exploration via randomized value functions [C]// International Conference on Machine

Learning. PMLR, 2016: 2377-2386.

[45] Zanette A, Brandfonbrener D, Brunskill E, *et al.* Frequentist regret bounds for randomized least-squares value iteration [C]// International Conference on Artificial Intelligence and Statistics. PMLR, 2020: 1954-1964.

[46] Azizzadenesheli K, Brunskill E, Anandkumar A. Efficient exploration through bayesian deep q-networks [C]// 2018 Information Theory and Applications Workshop (ITA) . IEEE, 2018: 1-9.

[47] Tang Y, Agrawal S. Exploration by distributional reinforcement learning [C]// Proceedings of the 27th International Joint Conference on Artificial Intelligence. 2018: 2710-2716.

[48] Colas C, Sigaud O, Oudeyer P Y. GEP-PG: Decoupling exploration and exploitation in deep reinforcement learning algorithms [C]// International conference on machine learning. PMLR, 2018: 1039-1048.

[49] Janz D, Hron J, Mazur P, *et al.* Successor uncertainties: exploration and uncertainty in temporal difference learning [J]. Advances in Neural Information Processing Systems, 2019, 32.

[50] Stulp F. Adaptive exploration for continual reinforcement learning [C]// 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2012: 1631-1636.

[51] Akiyama T, Hachiya H, Sugiyama M. Efficient exploration through active learning for value function approximation in reinforcement learning [J]. Neural Networks, 2010, 23 (5): 639-648.

[52] Strens M. A Bayesian framework for reinforcement learning [C]// ICML. 2000, 2000: 943-950.

[53] Guez A, Silver D, Dayan P. Efficient Bayes-adaptive reinforcement learning using sample-based search [J]. Advances in neural information processing systems, 2012, 25.

[54] Metelli A M, Likmeta A, Restelli M. Propagating uncertainty in reinforcement learning via wasserstein barycenters [J]. Advances in Neural Information Processing Systems, 2019, 32.

[55] O'Donoghue B, Osband I, Munos R, *et al.* The uncertainty bellman equation and exploration [C]// International Conference on Machine Learning. 2018: 3836-3845.

[56] Osband I, Blundell C, Pritzel A, *et al.* Deep exploration via bootstrapped DQN [J]. Advances in neural information processing systems, 2016, 29.

[57] Osband I, Aslanides J, Cassirer A. Randomized prior functions for deep reinforcement learning [J]. Advances in Neural Information Processing Systems, 2018, 31.

[58] Zhang Y, Goh W B. Bootstrapped policy gradient for difficulty adaptation in intelligent tutoring systems [C]// Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems. 2019: 711-719.

[59] Kalweit G, Boedecker J. Uncertainty-driven imagination for continuous deep reinforcement learning [C]// Conference on Robot Learning. PMLR, 2017: 195-206.

[60] Yang Z, Merrick K E, Abbass H A, *et al.* Multi-task deep reinforcement learning for continuous action control [C]// IJCAI. 2017, 17: 3301-3307.

[61] Zheng12 Z, Yuan C, Cheng12 Y. Self-adaptive double bootstrapped DDPG [C]// International Joint Conference on Artificial Intelligence. 2018.

[62] Peer O, Tessler C, Merlis N, *et al.* Ensemble bootstrapping for Q-Learning [C]// International Conference on Machine Learning. PMLR, 2021: 8454-8463.

[63] Hiraoka T, Imagawa T, Hashimoto T, *et al.* Dropout Q-functions for doubly efficient reinforcement learning [C]// International Conference on Learning Representations.

[64] Lee K, Laskin M, Srinivas A, *et al.* Sunrise: A simple unified framework for ensemble learning in deep reinforcement learning [C]// International Conference on Machine Learning. PMLR, 2021: 6131-6141.

[65] Pearce T, Anastassacos N, Zaki M, *et al.* Bayesian inference with anchored ensembles of neural networks, and application to exploration in reinforcement learning [EB/OL]. (2018-07-02) [2023-03-10] https://arxiv. org/pdf/1805. 11324. pdf.

[66] Ciosek K, Vuong Q, Loftin R, *et al.* Better exploration with optimistic actor critic [J]. Advances in Neural Information Processing Systems, 2019, 32.

[67] Bai C, Wang L, Han L, *et al.* Principled exploration via optimistic bootstrapping and backward induction [C]// International Conference on Machine Learning. PMLR, 2021: 577-587.

[68] Moerland T, Broekens J, Jonker C. Efficient exploration with Double Uncertain Value Networks [C]// NIPS 2017: Thirty-first Conference on Neural Information Processing Systems. 2017: 1-17.

[69] Nikolov N, Kirschner J, Berkenkamp F, *et al.* Information-directed exploration for deep reinforcement learning [C]// Online Program: 7th International Conference on Learning Representations (ICLR 2019) . ICRL, 2019.

[70] Kirschner J, Krause A. Information directed sampling and bandits with heteroscedastic noise [C]// Conference On Learning Theory. PMLR, 2018: 358-384.

[71] Bellemare M G, Dabney W, Munos R. A distributional perspective on reinforcement learning [C]// International conference on machine learning. PMLR, 2017: 449-458.

[72] Clements W R, Van Delft B, Robaglia B M, *et al.* Estimating risk and uncertainty in deep reinforcement learning [J]. arXiv preprint arXiv: 1905. 09638, 2019.

[73] Dabney W, Rowland M, Bellemare M, *et al.* Distributional reinforcement learning with quantile regression [C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2018, 32 (1) .

[74] Pearce T, Leibfried F, Brintrup A. Uncertainty in neural networks: Approximately bayesian ensembling [C]// International conference on artificial intelligence and statistics. PMLR, 2020: 234-244.

[75] Mavrin B, Yao H, Kong L, *et al.* Distributional reinforcement learning for efficient exploration [C]// International conference on machine learning. PMLR, 2019: 4424-4434.

[76] Zhou F, Wang J, Feng X. Non-crossing quantile regression for distributional reinforcement learning [J]. Advances in Neural Information Processing Systems, 2020, 33: 15909-15919.

[77] Mai V, Mani K, Paull L. Sample efficient deep reinforcement learning via uncertainty estimation [C]//International Conference on Learning Representations.

[78] Mavor-Parker A, Young K, Barry C, *et al.* How to stay curious while avoiding noisy tvs using aleatoric uncertainty estimation [C]// International Conference on Machine Learning. PMLR, 2022: 15220-15240.

[79] Kendall A, Gal Y. What uncertainties do we need in bayesian deep learning for computer vision? [J]. Advances in neural information processing systems, 2017, 30.

[80] Kulkarni T D, Narasimhan K, Saeedi A, *et al.* Hierarchical deep

reinforcement learning: Integrating temporal abstraction and intrinsic motivation [J]. Advances in neural information processing systems, 2016, 29.

[81] 陶卿, 马坡, 张梦晗, 等. 机器学习随机优化方法的个体收敛性研究综述 [J]. 数据采集与处理, 2017, 第 32 卷 (1): 17-25 (Tao Qing, Ma Po, Zhang Menghan, *et al.* Research on individual convergence of stochastic optimization methods for machine learning [J]. Data Acquisition and Processing, 2017, Vol 32 (1): 17-25)

[82] 纪守领, 杜天宇, 李进锋, 等. 机器学习模型安全与隐私研究综述 [J]. 软件学报, 2021, 32 (1): 41-67 (Ji Shouling, Du Tianyu, Li Jinfeng, *et al.* Survey on security and privacy of machine learning models [J]. Journal of Software, 2021, 32 (1): 41-6)

[83] Lipton Z C, Azizzadenesheli K, Kumar A, *et al.* Combating reinforcement learning's sisyphean curse with intrinsic fear [EB/OL]. (2018-03-13) [2023-03-10]. https://arxiv. org/pdf/1611. 01211. pdf.

[84] 杜威; 丁世飞. 多智能体强化学习综述 [J]. 计算机科学, 2019, 第 46 卷 (8): 1-8 (Du Wei; Ding Shifei. Survey on multi-agent reinforcement learning [J]. Computer Science, 2019, Vol. 46 (8): 1-8)

[85] Agarwal R, Schuurmans D, Norouzi M. An optimistic perspective on offline reinforcement learning [C]// International Conference on Machine Learning. PMLR, 2020: 104-114.

[86] Kumar A, Fu J, Soh M, *et al.* Stabilizing off-policy q-learning via bootstrapping error reduction [J]. Advances in Neural Information Processing Systems, 2019, 32.

[87] An G, Moon S, Kim J H, *et al.* Uncertainty-based offline reinforcement learning with diversified q-ensemble [J]. Advances in neural information processing systems, 2021, 34: 7436-7447.

[88] Haarnoja T, Zhou A, Abbeel P, *et al.* Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor [C]// International conference on machine learning. PMLR, 2018: 1861-1870.

[89] An G, Moon S, Kim J H, *et al.* Uncertainty-based offline reinforcement learning with diversified q-ensemble [J]. Advances in neural information processing systems, 2021, 34: 7436-7447.

[90] Lee S, Seo Y, Lee K, *et al.* Offline-to-online reinforcement learning via balanced replay and pessimistic q-ensemble [C]// Conference on Robot Learning. PMLR, 2022: 1702-1712.