# 安全强化学习综述

王雪松 [1] 王荣荣 [1] 程玉虎 [1]

**摘 要** 强化学习 (Reinforcement learning, RL) 在围棋、视频游戏、导航、推荐系统等领域均取得了巨大成功. 然而, 许多强化学习算法仍然无法直接移植到真实物理环境中. 这是因为在模拟场景下智能体能以不断试错的方式与环境进行交互, 从而学习最优策略. 但考虑到安全因素, 很多现实世界的应用则要求限制智能体的随机探索行为. 因此, 安全问题成为强化学习从模拟到现实的一个重要挑战. 近年来, 许多研究致力于开发安全强化学习 (Safe reinforcement learning, SRL) 算法, 在确保系统性能的同时满足安全约束. 本文对现有的安全强化学习算法进行全面综述, 将其归为三类: 修改学习过程、修改学习目标、离线强化学习, 并介绍了 5 大基准测试平台: Safety Gym、safe-control-gym、SafeRL-Kit、D4RL、NeoRL. 最后总结了安全强化学习在自动驾驶、机器人控制、工业过程控制、电力系统优化和医疗健康领域中的应用, 并给出结论与展望.

**关键词** 安全强化学习, 约束马尔科夫决策过程, 学习过程, 学习目标, 离线强化学习

**引用格式** 王雪松, 王荣荣, 程玉虎. 安全强化学习综述. 自动化学报, 2023, **49**(9): 1813−1835

**DOI** 10.16383/j.aas.c220631

## Safe Reinforcement Learning: A Survey

WANG Xue-Song[1] WANG Rong-Rong[1] CHENG Yu-Hu[1]

**Abstract** Reinforcement learning (RL) has proved a prominent success in the game of Go, video games, navigation, recommendation systems and other fields. However, a large number of reinforcement learning algorithms cannot be directly transplanted to real physical environment. This is because in the simulation scenario, the agent is able to interact with the environment in a trial-and-error manner to learn the optimal policy. Considering the safety of systems, many real-world applications require the limitation of random exploration behavior of agents. Hence, safety has become an essential factor for reinforcement learning from simulation to reality. In recent years, many researches have been devoted to develope safe reinforcement learning (SRL) algorithms that satisfy safety constraints while ensuring system performance. This paper presents a comprehensive survey of existing SRL algorithms, which are divided into three categories: Modification of learning process, modification of learning objective, and offline reinforcement learning. Furthermore, five experimental platforms are introduced, including Safety Gym, safe-control-gym, SafeRL-Kit, D4RL, and NeoRL. Lastly, the applications of SRL in the fields of autonomous driving, robot control, industrial process control, power system optimization, and healthcare are summarized, and the conclusion and perspective are briefly drawn.

**Key words** Safe reinforcement learning (SRL), constrained Markov decision process (CMDP), learning process, learning objective, offline reinforcement learning

**Citation** Wang Xue-Song, Wang Rong-Rong, Cheng Yu-Hu. Safe reinforcement learning: A survey. *Acta Automatica Sinica*, 2023, **49**(9): 1813−1835

作为一种重要的机器学习方法, 强化学习 (Reinforcement learning, RL) 采用了人类和动物学习中 “试错法” 与 “奖惩回报” 的行为心理学机制, 强调智能体在与环境的交互中学习, 利用评价性的反馈信号实现决策的优化[1]. 早期的强化学习主要依赖于人工提取特征, 难以处理复杂高维状态和动作空间下的问题. 近年来, 随着计算机硬件设备性能的提升和神经网络学习算法的发展, 深度学习由于其强大的表征能力和泛化性能受到了众多研究人员的关注[2−3]. 于是, 将深度学习与强化学习相结合就成为了解决复杂环境下感知决策问题的一个可行方案. 2016 年, Google 公司的研究团队 DeepMind 创新性地将具有感知能力的深度学习与具有决策能

收稿日期 2022-08-08 录用日期 2023-01-11
Manuscript received August 8, 2022; accepted January 11, 2023
国家自然科学基金 (62176259, 61976215), 江苏省重点研发计划项目 (BE2022095) 资助
Supported by National Natural Science Foundation of China (62176259, 61976215) and Key Research and Development Program of Jiangsu Province (BE2022095)
本文责任编委 黎铭
Recommended by Associate Editor LI Ming
1. 中国矿业大学信息与控制工程学院 徐州 221116
1. School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116

力的强化学习相结合, 开发的人工智能机器人 AlphaGo 成功击败了世界围棋冠军李世石[4], 一举掀起了深度强化学习的研究热潮. 目前, 深度强化学习在视频游戏[5]、自动驾驶[6]、机器人控制[7]、电力系统优化[8]、医疗健康[9] 等领域均得到了广泛的应用.

近年来, 学术界与工业界开始逐步注重深度强化学习如何从理论研究迈向实际应用. 然而, 要实现这一阶段性的跨越还有很多工作需要完成, 其中尤为重要的一项任务就是保证决策的安全性. 安全对于许多应用至关重要, 一旦学习策略失败则可能会引发巨大灾难. 例如, 在医疗健康领域, 微创手术机器人辅助医生完成关于大脑或心脏等关键器官手术时, 必须做到精准无误, 一旦偏离原计划位置, 则将对病人造成致命危害. 再如, 自动驾驶领域, 如果智能驾驶车辆无法规避危险路障信息, 严重的话将造成车毁人亡. 因此, 不仅要关注期望回报最大化, 同时也应注重学习的安全性.

García 和 Fernández[10] 于 2015 年给出了安全强化学习 (Safe reinforcement learning, SRL) 的定义: 考虑安全或风险等概念的强化学习. 具体而言, 所谓安全强化学习是指在学习或部署过程中, 在保证合理性能的同时满足一定安全约束的最大化长期回报的强化学习过程. 自 2015 年起, 基于此研究, 学者们提出了大量安全强化学习算法. 为此, 本文对近年来的安全强化学习进行全面综述, 围绕智能体的安全性问题, 从修改学习过程、修改学习目标以及离线强化学习三方面进行总结, 并给出了用于安全强化学习的 5 大基准测试平台: Safety Gym、safe-control-gym、SafeRL-Kit、D4RL、NeoRL, 以及安全强化学习在自动驾驶、机器人控制、工业过程控制、电力系统优化以及医疗健康领域的应用. 安全强化学习中所涉及的方法、基准测试平台以及应用领域之间的关系如图 1 所示.

本文结构如下: 第 1 节对安全强化学习问题进行形式化描述; 第 2 节对近年来的安全强化学习方法进行分类与综述; 第 3 节介绍 5 种基准测试平台; 第 4 节总结安全强化学习的实际应用场景; 第 5 节对未来研究方向进行探讨; 第 6 节对文章进行总结.

## 1 问题描述

安全强化学习问题通常被定义为一个约束马尔科夫决策过程 (Constrained Markov decision process, CMDP) $\mathcal{M} \cup \mathcal{C}$ [11], 即在标准马尔科夫决策过程 $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \gamma, r \rangle$ 的基础上添加了关于成本函数的约束项 $\mathcal{C} = \{c, d\}$. $\mathcal{S}$ 表示状态空间集, $\mathcal{A}$ 表示动作空间集, $\mathcal{T}(s'|s, a)$ 表示用于描述动力学模型的状态转移函数, $\gamma$ 表示折扣因子, $r: \mathcal{S} \times \mathcal{A} \to \mathbf{R}$ 表示奖励函数; $c: \mathcal{S} \times \mathcal{A} \to \mathbf{R}$ 表示成本函数, $d$ 表示安全阈值. 这种情况下, 安全强化学习问题可以表述为在满足安全约束的情况下, 求解使期望回报最大化的最优可行策略 $\pi^*$

$$\pi^* = \arg\max_{\pi \in \Pi_c} J(\pi) \tag{1}$$

其中, $J(\pi) = \mathrm{E}_{\tau \sim \pi}\left(\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right)$, $\tau = (s_0, a_0, s_1, a_1, \cdots)$ 表示一条轨迹, $\tau \sim \pi$ 表示轨迹 $\tau$ 根据策略 $\pi$ 采样得到, $\Pi_c$ 表示满足安全约束的安全策略集. 值得注意的是, 本文公式所描述的都是单成本约束的形式, 但不失一般性, 这些公式都可以拓展为多成本约束的形式. 对于不同类型的决策任务, 安全策略集可以有不同的表达形式.

对于安全性要求严格的决策任务, 例如自动驾驶[12−13] 任务, 通常采用硬约束方式, 即在所有的时刻都需要强制满足单步约束. 这种情况下 $\Pi_c$ 表示为

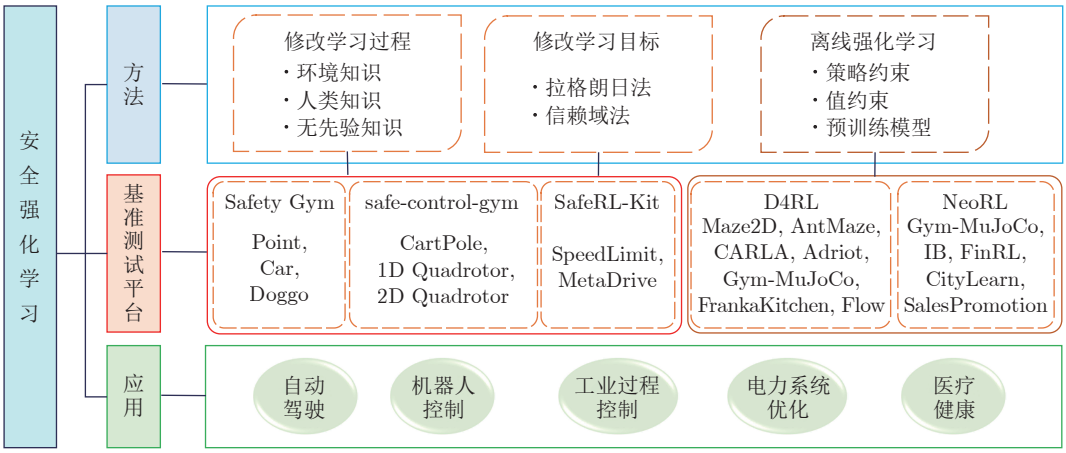



图 1 安全强化学习方法、基准测试平台与应用

Fig. 1 Methods, benchmarking platforms, and applications of safe reinforcement learning

$$\Pi_c = \{\pi \in \Pi\colon\ c(s_t,\, a_t) < d,\, \forall t \in \{0,\, 1,\, \cdots\}\} \quad (2)$$

其中, $\Pi$ 表示可行策略集. 但由于这种约束方式要求过于严格, 因此通常需要借助模型信息加以实现.

在无模型情况下, 软约束方式有着更广泛的应用, 即对折扣累积成本的期望进行约束, 这种情况下 $\Pi_c$ 表示为

$$\Pi_c = \left\{\pi \in \Pi\colon \mathop{\mathrm{E}}_{\tau\sim\pi}\left(\sum_{t=0}^{\infty}\gamma^t c(s_t,\, a_t)\right) < d\right\} \quad (3)$$

这种约束方式可以很好地适用于机器人行走[14]、油泵安全控制[15] 和电力系统优化[16] 等任务, 但对于需要明确定义状态或动作是否安全的任务却难以处理. 为了使软约束方式更好地适用于不同类型的决策任务, 可以将成本函数修改为 $c\colon \mathcal{S}\times\mathcal{A}\to\{0,\,1\}$, 利用成本函数对当前状态动作对进行安全性判断, 若安全, 则 $c(s_t,\, a_t) = 0$, 否则, $c(s_t,\, a_t) = 1$, 并且在智能体与环境交互期间遇到不安全的状态动作对时终止当前回合. 这时, 约束项 $\mathrm{E}_{\tau\sim\pi}(\sum_{t=0}^{\infty}\gamma^t c(s_t,\, a_t))$ 可以表示 $\pi$ 产生不安全状态动作对的概率, 因此经过这样修改后的软约束也被称为机会型约束. 机会型约束由于其良好的任务适应性, 已被成功应用于无模型的自动驾驶[17] 和机械臂控制[18] 等任务.

另一方面, 离线强化学习[19−20] 从一个静态的数据集中学习最优策略, 它避免了与环境的交互过程, 可以保障训练过程中的安全性. 因此, 可以将离线强化学习作为安全强化学习的一种特殊形式. 离线强化学习考虑一个标准马尔科夫决策过程 $\mathcal{M} = \langle\mathcal{S},\, \mathcal{A},\, \mathcal{T},\, \gamma,\, r\rangle$, 它的目标是求解使期望回报最大化的最优可行策略 $\pi^* = \arg\max_{\pi\in\Pi} J(\pi)$, 与在线方式不同的是, 智能体在训练过程中不再被允许与环境进行交互, 而是只能从一个静态数据集 $\mathcal{B} = \{(s,\, a,\, r,\, s')\}$ 中进行学习. 尽管这种方式可以保障训练过程中的安全性, 但分布偏移问题 (目标策略与行为策略分布不同)[19−20] 也给求解 $\pi^*$ 的过程带来了困难. 因此, 现如今的离线强化学习方法大多关注于如何解决分布偏移问题. 离线强化学习在有先验离线数据集支持的情况下, 借助于其训练过程安全的优势, 已被应用于微创手术机器人控制[21] 和火力发电机组控制[22] 等任务.

## 2 方法分类

求解安全强化学习问题的方法有很多, 受 García 和 Fernández[10] 启发, 本文从以下三方面进行综述:

1) 修改学习过程. 通过约束智能体的探索范围, 采用在线交互反馈机制, 在强化学习的学习或探索过程中阻止其产生危险动作, 从而确保了训练时策略的安全性. 根据是否利用先验知识, 将此类方法划分为三类: 环境知识、人类知识、无先验知识.

2) 修改学习目标. 同样采用在线交互反馈机制, 在强化学习的奖励函数或目标函数中引入风险相关因素, 将约束优化问题转化为无约束优化问题, 如拉格朗日法、信赖域法.

3) 离线强化学习. 仅在静态的离线数据集上训练而不与环境产生交互, 从而完全避免了探索, 但对部署时安全没有任何约束保证, 并未考虑风险相关因素. 因此大多数离线强化学习能实现训练时安全, 但无法做到部署时安全.

三类安全强化学习方法的适用条件、优缺点以及应用领域对比如表 1 所示. 下面对安全强化学习的现有研究成果进行详细综述与总结.

### 2.1 修改学习过程

在强化学习领域, 智能体需要通过不断探索来减小外界环境不确定性对自身学习带来的影响. 因此, 鼓励智能体探索一直是强化学习领域非常重要的一个研究方向. 然而, 不加限制的自由探索很有可能使智能体陷入非常危险的境地, 甚至酿成重大安全事故. 为避免强化学习智能体出现意外和不可逆的后果, 有必要在训练或部署的过程中对其进行安全性评估并将其限制在 “安全” 的区域内进行探索, 将此类方法归结为修改学习过程. 根据智能体利用先验知识的类型将此类方法进一步细分为环境知识、人类知识以及无先验知识. 其中环境知识利用系统动力学先验知识实现安全探索; 人类知识借鉴人类经验来引导智能体进行安全探索; 无先验知识没有用到环境知识和人类知识, 而是利用安全约束结构将不安全的行为转换到安全状态空间中.

#### 2.1.1 环境知识

基于模型的方法因其采样效率高而得以广泛研究. 该类方法利用了环境知识, 需要学习系统动力学模型, 并利用模型生成的轨迹来增强策略学习, 其核心思想就是通过协调模型使用和约束策略搜索来提高安全探索的采样效率. 可以使用高斯过程对模型进行不确定性估计, 利用 Shielding 修改策略动作从而生成满足约束的安全过滤器, 使用李雅普诺夫函数法或控制障碍函数法来限制智能体的动作选择, 亦或使用已学到的动力学模型预测失败并生成安全策略. 具体方法总结如下.

高斯过程. 一种主流的修改学习过程方式是使用高斯过程对具有确定性转移函数和值函数的动力

表 1 安全强化学习方法对比

Table 1 Comparison of safe reinforcement learning methods

| 方法类别 | | 训练时安全 | 部署时安全 | 与环境实时交互 | 优点 | 缺点 | 应用领域 |
|---|---|---|---|---|---|---|---|
| 修改学习过程 | 环境知识 | √ | √ | √ | 采样效率高 | 需获取环境的动力学模型、实现复杂 | 自动驾驶[12−13, 23]、工业过程控制[24−25]、电力系统优化[26]、医疗健康[21] |
| | 人类知识 | √ | √ | √ | 加快学习过程 | 人工监督成本高 | 机器人控制[14, 27]、电力系统优化[28]、医疗健康[29] |
| | 无先验知识 | √ | √ | √ | 无需获取先验知识、可扩展性强 | 收敛性差、训练不稳定 | 自动驾驶[30]、机器人控制[31]、工业过程控制[32]、电力系统优化[33]、医疗健康[34] |
| 修改学习目标 | 拉格朗日法 | × | √ | √ | 思路简单、易于实现 | 拉格朗日乘子选取困难 | 工业过程控制[15]、电力系统优化[16] |
| | 信赖域法 | √ | √ | √ | 收敛性好、训练稳定 | 近似误差不可忽略、采样效率低 | 机器人控制[35] |
| 离线强化学习 | 策略约束 | √ | × | × | 收敛性好 | 方差大、采样效率低 | 医疗健康[36] |
| | 值约束 | √ | × | × | 值函数估计方差小 | 收敛性差 | 工业过程控制[22] |
| | 预训练模型 | √ | × | × | 加快学习过程、泛化性强 | 实现复杂 | 工业过程控制[37] |

学建模, 以便能够估计约束和保证安全学习. Sui 等[38] 将 “安全” 定义为: 在智能体学习过程中, 选择的动作所收到的期望回报高于一个事先定义的阈值. 由于智能体只能观测到当前状态的安全函数值, 而无法获取相邻状态的信息, 因此需要对安全函数进行假设. 为此, 在假设回报函数满足正则性、Lipschitz 连续以及范数有界等条件的前提下, Sui 等[38] 利用高斯过程对带参数的回报函数进行建模, 提出一种基于高斯过程的安全探索方法 SafeOpt. 在学习过程中, 结合概率生成模型, 通过贝叶斯推理即可求得高斯过程的后验分布, 即回报函数空间的后验. 进一步, 利用回报函数置信区间来评估决策的安全性, 得到一个安全的参数区间并约束智能体只在这个安全区间内进行探索. 然而, SafeOpt 仅适用于类似多臂老虎机这类的单步、低维决策问题, 很难推广至复杂决策问题. 为此, Turchetta 等[39] 利用马尔科夫决策过程的可达性, 在 SafeOpt 的基础上提出 SafeMDP 安全探索方法, 使其能够解决确定性有限马尔科夫决策过程问题. 在 SafeOpt 和 SafeMDP 中, 回报函数均被视为是先验已知和时不变的, 但在很多实际问题中, 回报函数通常是先验未知和时变的. 因此, 该方法并未在考虑安全的同时优化回报函数. 针对上述问题, Wachi 等[40] 把时间和空间信息融入核函数, 利用时−空高斯过程对带参数的回报函数进行建模, 提出一种新颖的安全探索方法: 时−空 SafeMDP (Spatio-temporal SafeMDP, ST-SafeMDP), 能够依概率确保安全性并同时优化回报目标. 尽管上述方法是近似安全的, 但正则性、Lipschitz 连续以及范数有界这些较为严格的假设条件限制了 SafeOpt、SafeMDP 和 ST-SafeMDP 在实际中的应用, 而且, 此类方法存在理论保证与计算成本不一致的问题, 在高维空间中很难达到理论上保证的性能.

Shielding. Alshiekh 等[41] 首次提出 Shielding 的概念来确保智能体在学习期间和学习后保持安全. 根据 Shielding 在强化学习环节中部署的位置, 将其分为两种类型: 前置 Shielding 和后置 Shielding. 前置 Shielding 是指在训练过程中的每个时间步, Shielding 仅向智能体提供安全的动作以供选择. 后置 Shielding 方式较为常用, 它主要影响智能体与环境的交互过程, 如果当前策略不安全则触发 Shielding, 使用一个备用策略来覆盖当前策略以保证安全性. 可以看出, 后置 Shielding 方法的使用主要涉及两个方面的工作: 1) Shielding 触发条件的设计. Zhang 等[42] 通过一个闭环动力学模型来估计当前策略下智能体未来的状态是否为可恢复状态, 如果不可恢复, 则需要采用备用策略将智能体还原到初始状态后再重新训练. 但如果智能体的状态不能还原, 则此方法就会失效. Jansen 等[43] 一方面采用形式化验证的方法来计算马尔科夫决策过程安全片段中关键决策的概率, 另一方面根据下一步状态的安全程度来估计决策的置信度. 当关键决策的概率及其置信度均较低时, 则启用备用策略. 但是, 在复杂的强化学习任务中, 从未知的环境中提取出安全片段并不是一件容易的事情. 2) 备用 (安全) 策略的设计. Li 和 Bastani[44] 提出了一种基于 tube 的鲁棒非线性模型预测控制器并将其作为备用控制器, 其中 tube 为某策略下智能体多次运行轨迹组成的集合. Bastani[45] 进一步将备用策略划分为不变策略和恢复策略, 其中不变策略使智能体在安全平衡点附近运动, 恢复策略使智能体运行到安全平衡点. Shielding 根据智能体与安全平衡点的距离来

决定选用何种类型的备用策略, 从而进一步增强了智能体的安全性. 但是, 在复杂的学习问题中, 很难定义安全平衡点, 往往也无法直观地观测状态到平衡点的距离. 综上所述, 如果环境中不存在可恢复状态, Shielding 即便判断出了危险, 也没有适合的备用策略可供使用. 此外, 在复杂的强化学习任务中, 很难提供充足的先验知识来搭建一个全面的 Shielding 以规避所有的危险.

李雅普诺夫法. 李雅普诺夫稳定性理论对于控制理论学科的发展产生了深刻的影响, 是现代控制理论中一个非常重要的组成部分. 该方法已被广泛应用于控制工程中以设计出达到定性目标的控制器, 例如稳定系统或将系统状态维持在所需的工作范围内. 李雅普诺夫函数可以用来解决约束马尔科夫决策过程问题并保证学习过程中的安全性. Perkins 和 Barto[46] 率先提出了在强化学习中使用李雅普诺夫函数的思路, 通过定性控制技术设计一些基准控制器并使智能体在这些给定的基准控制器间切换, 用于保证智能体的闭环稳定性. 为了规避风险, 要求强化学习方法具有从探索动作中安全恢复的能力, 也就是说, 希望智能体能够恢复到安全状态. 众所周知, 这种状态恢复的能力就是控制理论中的渐近稳定性. Berkenkamp 等[47] 使用李雅普诺夫函数对探索空间进行限制, 让智能体大概率地探索到稳定的策略, 从而能够确保基于模型的强化学习智能体可以在探索过程中被带回到 “吸引区域”. 所谓吸引区域是指: 状态空间的子集, 从该集合中任一状态出发的状态轨迹始终保持在其中并最终收敛到目标状态. 然而, 该方法只有在满足 Lipschitz 连续性假设条件下才能逐步探索安全状态区域, 这需要事先对具体系统有足够了解, 一般的神经网络可能并不具备 Lipschitz 连续. 上述方法是基于值函数的, 因此将其应用于连续动作问题上仍然具有挑战性. 相比之下, Chow 等[48] 更专注于策略梯度类方法, 从原始 CMDP 安全约束中生成一组状态相关的李雅普诺夫约束, 提出一种基于李雅普诺夫函数的 CMDP 安全策略优化方法. 主要思路为: 使用深度确定性策略梯度和近端策略优化算法训练神经网络策略, 同时通过将策略参数或动作映射到由线性化李雅普诺夫约束诱导的可行解集上来确保每次策略更新时的约束满意度. 所提方法可扩展性强, 能够与任何同策略或异策略的方法相结合, 可以处理具有连续动作空间的问题, 并在训练和收敛过程中返回安全策略. 通过使用李雅普诺夫函数和 Transformer 模型, Jeddi 等[49] 提出一种新的不确定性感知的安全强化学习算法. 该算法主要思路为: 利用具有理论安全保证的李雅普诺夫函数将基于轨迹的安全约束转换为一组基于状态的局部线性约束; 将安全强化学习模型与基于 Transformer 的编码器模型相结合, 通过自注意机制为智能体提供处理长时域范围内信息的记忆; 引入一个规避风险的动作选择方案, 通过估计违反约束的概率来识别风险规避的动作, 从而确保动作的安全性. 总而言之, 李雅普诺夫方法的主要特征是将基于轨迹的约束分解为一系列单步状态相关的约束. 因此, 当状态空间无穷大时, 可行性集就具有无穷维约束的特征, 此时直接将这些李雅普诺夫约束 (相对于原始的基于轨迹的约束) 强加到策略更新优化中实现成本高, 无法应用于真实场景, 而且, 此类方法仅适用于基于模型的强化学习且李雅普诺夫函数通常难以构造.

障碍函数法. 障碍函数法是另一种保证控制系统安全的方法. 其基本思想为: 系统状态总是从内点出发, 并始终保持在可行安全域内搜索. 在原先的目标函数中加入障碍函数惩罚项, 相当于在可行安全域边界构筑起一道 “墙”. 当系统状态达到安全边界时, 所构造的障碍函数值就会趋于无穷, 从而避免状态处于安全边界, 而是被 “挡” 在安全域内. 为保证强化学习算法在模型信息不确定的情况下的安全性, Cheng 等[50] 提出了一种将现有的无模型强化学习算法与控制障碍函数 (Control barrier functions, CBF) 相结合的框架 RL-CBF. 该框架利用高斯过程来模拟系统动力学及其不确定性, 通过使用预先指定的障碍函数来指导策略探索, 提高了学习效率, 实现了非线性控制系统的端到端安全强化学习. 然而, 使用的离散时间 CBF 公式具有限制性, 因为它只能通过仿射 CBF 的二次规划进行实时控制综合. 例如, 在避免碰撞的情况下, 仿射 CBF 只能编码多面体障碍物. 为了在学习过程中保持安全性, 系统状态必须始终保持在安全集内, 该框架前提假设已得到一个有效安全集, 但实际上学习安全集并非易事, 学习不好则可能出现不安全状态. Yang 等[51] 采用障碍函数对系统进行变换, 将原问题转化为无约束优化问题的同时施加状态约束. 为减轻通信负担, 设计了静态和动态两类间歇性策略. 最后, 基于 actor-critic 架构, 提出一种安全的强化学习算法, 采用经验回放技术, 利用历史数据和当前数据来共同学习约束问题的解, 在保证最优性、稳定性和安全性的同时以在线的方式寻求最优安全控制器. Marvi 和 Kiumarsi[52] 提出了一种安全异策略强化学习方法, 以数据驱动的方式学习最优安全策略. 该方法将 CBF 合并进安全最优控制成本目标中形成一个增广值函数, 通过对该增广值函数进行迭代近似并调节权衡因子, 从而实现安全性与最优性的平衡. 但在实际应用中, 权衡因子的选取需要事先

人工设定, 选择不恰当则可能找不到最优解. 先前的工作集中在一类有限的障碍函数上, 并利用一个辅助神经网来考虑安全层的影响, 这本身就造成了一种近似. 为此, Emam 等[53] 将一个可微的鲁棒控制障碍函数 (Robust CBF, RCBF) 层合并进基于模型的强化学习框架中. 其中, RCBF 可用于非仿射实时控制综合, 而且可以对动力学上的各种扰动进行编码. 同时, 使用高斯过程来学习扰动, 在安全层利用扰动生成模型轨迹. 实验表明, 所提方法能有效指导训练期间的安全探索, 提高样本效率和稳态性能. 障碍函数法能够确保系统安全, 但并未考虑系统的渐进稳定性, 与李雅普诺夫法类似, 在实际应用中障碍函数和权衡参数都需要精心设计与选择.

引入惩罚项. 此类方法在原先目标函数的基础上添加惩罚项, 以此修正不安全状态. 由于传统的乐观探索方法可能会使智能体选择不安全的策略, 导致违反安全约束, 为此, Bura 等[54] 提出一种基于模型的乐观–悲观安全强化学习算法 (Optimistic-pessimistic SRL, OPSRL). 该算法在不确定性乐观目标函数的基础上添加悲观约束成本函数惩罚项, 对回报目标持乐观态度以便促进探索, 同时对成本函数持悲观态度以确保安全性. 在 Media Control 环境下的仿真结果表明, OPSRL 在没有违反安全约束的前提下能获得最优性能. 基于模型的方法有可能在安全违规行为发生之前就得以预测, 基于这一动机, Thomas 等[55] 提出了基于模型的安全策略优化算法 (Safe model-based policy optimization, SMBPO). 该算法通过预测未来几步的轨迹并修改奖励函数来训练安全策略, 对不安全的轨迹进行严厉惩罚, 从而避免不安全状态. 在 MuJoCo 机器人控制模拟环境下的仿真结果表明, SMBPO 能够有效减少连续控制任务的安全违规次数. 但是, 需要有足够大的惩罚和精确的动力学模型才能避免违反安全. Ma 等[56] 提出了一种基于模型的安全强化学习方法, 称为保守与自适应惩罚 (Conservative and adaptive penalty, CAP). 该方法使用不确定性估计作为保守惩罚函数来避免到达不安全区域, 确保所有的中间策略都是安全的, 并在训练过程中使用环境的真实成本反馈适应性地调整这个惩罚项, 确保零安全违规. 相比于先前的安全强化学习算法, CAP 具有高效的采样效率, 同时产生了较少的违规行为.

### 2.1.2 人类知识

为了获得更多的经验样本以充分训练深度网络, 有些深度强化学习方法甚至在学习过程中特意加入带有随机性质的探索性学习以增强智能体的探索能力. 一般来说, 这种自主探索仅适用于本质安全的系统或模拟器. 如果在现实世界的一些任务 (例如智能交通、自动驾驶) 中直接应用常规的深度强化学习方法, 让智能体进行不受任何安全约束的 “试错式” 探索学习, 所做出的决策就有可能使智能体陷入非常危险的境地, 甚至酿成重大安全事故. 相较于通过随机探索得到的经验, 人类专家经验具备更强的安全性. 因此, 借鉴人类经验来引导智能体进行探索是一个可行的增强智能体安全性的措施. 常用的方法有中断机制、结构化语言约束、专家指导.

中断机制. 此类方法借鉴了人类经验, 当智能体做出危险动作时能及时进行中断. 在将强化学习方法应用于实际问题时, 最理想的状况是智能体任何时候都不会做出危险动作. 由于限制条件太强, 只能采取 “人在环中” 的人工介入方式, 即人工盯着智能体, 当出现危险动作时, 出手中断并改为安全的动作. 但是, 让人来持续不断地监督智能体进行训练是不现实的, 因此有必要将人工监督自动化. 基于这个出发点, Saunders 等[57] 利用模仿学习技术来学习人类的干预行为, 提出一种人工干预安全强化学习 (SRL via human intervention, HIRL) 方法. 主要思路为: 首先, 在人工监督阶段, 收集每一个状态–动作对以及与之对应的 “是否实施人工中断” 的二值标签; 然后, 基于人工监督阶段收集的数据, 采用监督学习方式训练一个 “Blocker” 以模仿人类的中断操作. 需要指出的是, 直到 “Blocker” 在剩余的训练数据集上表现良好, 人工监督阶段的操作方可停止. 采用 4 个 Atari 游戏来测试 HIRL 的性能, 结果发现: HIRL 的应用场景非常受限, 仅能处理一些较为简单的智能体安全事故且难以保证智能体完全不会做出危险动作; 当环境较为复杂的时候, 甚至需要一年以上的时间来实施人工监督, 时间成本高昂. 为降低时间成本, Prakash 等[58] 将基于模型的方法与 HIRL 相结合, 提出一种混合安全强化学习框架, 主要包括三个模块: 基于模型的模块、自举模块、无模型模块. 首先, 基于模型的模块由一个动力学模型组成, 用以驱动模型预测控制器来防止危险动作发生; 然后, 自举模块采用由模型预测控制器生成的高质量示例来初始化无模型强化学习方法的策略; 最后, 无模型模块使用基于自举策略梯度的强化学习智能体在 “Blocker” 的监督下继续学习任务. 但是, 作者仅在小规模的 4×4 格子世界和 Island Navigation 仿真环境中验证了方法的有效性, 与 HIRL 一样, 该方法的应用场景仍

然非常受限. 前两种方法都是通过 “Blocker” 模仿人类干预来识别危险动作. 但通常情况下, 人类在遇到任何潜在的危险时, 会立即停止行动. 受此启发, Sun 等[59] 提出一种解决提前终止马尔科夫决策过程 (Early terminated Markov decision process, ET-MDP) 框架下安全强化学习问题的方法. 该方法以最直接的方式处理约束条件: 只要学习策略违反约束条件, 就会触发提前终止, 从而避免不安全行为. 进一步引入了基于上下文模型的异策略算法来缓解 ET-MDP 求解中的有限状态访问问题. 在一系列 CMDP 环境中评估了所提方法, 结果表明在约束条件下 ET-MDP 的学习效率和渐近性能均有显著提高. 但在某些约束严格的场景下, 使用提前终止原则可能会失效, 导致无法对约束马尔科夫决策问题进行求解.

结构化语言约束. 增强强化学习智能体安全性的另一种方法是借助人类的自然语言表述将抽象的安全标准映射为成本函数或回报函数. 在深度强化学习中, 有很多任务对应的回报是比较稀疏的, 不仅会使学习效率降低, 而且易于生成次优策略和导致不安全的行为. 为此, Prakash 等[60] 设计了一种以专家结构化语言生成约束条件的安全深度强化学习框架. 主要思路为: 首先, 收集由轨迹片段和结构化语言约束构成的数据集; 然后, 根据是否发生违规行为对数据进行人工手动标记; 最后, 利用该数据集来训练 “约束检查器”, 这是一种用于在智能体训练期间提供辅助回报信号的嵌入模型. 约束检查器用以判断智能体的动作是否与专家意见相冲突, 如果发生冲突则降低相应的回报值. 在后续的训练过程当中, 智能体就更倾向于减小与专家意见的冲突, 提高动作的安全性. 这项工作的局限性在于: 它需要一种方法来自动生成数据以训练约束检查器, 或者需要人工来收集和标记数据. 先前的安全强化学习方法大都以数学或逻辑形式表达约束, 因此需要具备特定的领域专业知识, 而且使用特定安全约束集训练的策略不易迁移到使用相同约束集的新任务中, 面对新任务仍需要从头训练. 为此, Yang 等[61] 通过手工定义的成本函数来强制执行安全标准, 提出一种基于语言约束的策略优化 (Policy optimization with language constraints, POLCO) 方法. 主要包括两个部分: 约束解释器, 将自然语言约束编码为中间向量或矩阵表征形式, 用以捕获禁止状态的时空信息; 策略网络, 根据约束解释器输出的表征形式, 利用约束策略优化输出具有最小约束违规的策略. 利用自然语言能够轻松灵活地指定安全约束, 且可以对模块化约束解释器进行重用从而适应新任务, 但当将这些约束映射到用于安全强化学习的表征形式时, 其模棱两可的性质带来了较大的挑战, 而且, 与文献 [60] 类似, 该方法需要为约束解释器设置显式标签. 如果将此类方法扩展到更为复杂的环境或现实世界中, 自动管理此类数据集将不可行, 而必须由人工来手动收集, 不仅费时耗力, 而且很难避免数据的错采、漏采、错标以及漏标.

专家指导. 专家指导信息可以使得智能体更快地学习到与专家策略相近的策略, 从而提高学习的安全性. 课程学习的主要思想是模仿人类学习的特点, 由简单到困难来学习课程 (在机器学习里就是容易学习的样本和不容易学习的样本), 从而有助于模型找到更好的局部最优, 同时加快训练的速度. Turchetta 等[62] 首次将课程学习引入到安全强化学习场景中, 提出一种受人类教学启发的安全强化学习方法. 智能体 (学生) 在监督者 (教师) 的自动指导下进行学习, 监督者会在智能体开始出现危险行为时选择不同的重置/干预动作, 从而避免智能体在学习过程中违反约束. 监督者根据智能体的学习进度和行为数据分布, 训练一个决策模型来自动选择重置/干预动作类型, 从而对智能体的课程学习进行自动设计. 然而, 该方法将安全探索全权交给监督者, 而监督者的好坏需要根据实际数据训练得到, 如何减小监督者的样本复杂度是具有挑战性的问题. 人类初学者在学习驾驶时, 通常会有领域专家 (监护人) 在旁边对其指导以确保学习过程中的安全性, 避免危险事故的发生. 基于该想法, Peng 等[63] 提出一种基于专家指导的策略优化 (Expert guided policy optimization, EGPO) 方法. 该方法在智能体与环境的交互中加入了监护人机制, 通过基于近端策略优化 (Proximal policy optimization, PPO) 的专家策略来监控学习智能体, 进一步引入离线强化学习技术, 从监护人那里收集到部分示范数据以便指导学习. 与先前方法相比, EGPO 具有更好的泛化性能和采样效率. 但过多的人工干预会增加成本, 为此, Li 等[64] 提出了一种高效的人−人工智能副驾驶优化 (Human-AI copilot optimization, HACO) 方法, 将人类融入到智能体与环境的交互中, 保证了安全高效的探索. 通过离线强化学习技术将由部分示范中提取的人类知识注入代理值函数, 利用最大熵正则化鼓励人类对状态动作空间的探索, 同时, 最小化训练过程中的人为干预成本, 减小随时间推移对专家示范的依赖, 提高智能体学习的自主性. 与 EGPO 的不同之处在于, 所提方法进一步减少了人类专家干预, 设计专门的机制来减小延迟反馈误差. 在实验设计方面, 去除了奖励函

数等冗余设计, 使方法更简单有效. 实验表明, 该方法具有较高的采样效率和安全性保证. 但与强化学习基准算法相比, HACO 在训练后表现过于保守, 智能体倾向于缓慢驾驶, 在十字路口出现频繁让路等行为. 综上所述, 专家指导类的方法需要人工干预, 监督智能体的行为, 阻止其探索危险状态. 这种方法能同时做到训练时安全和部署时安全, 但在实际应用中, 人工干预的成本也需要考虑在内.

#### 2.1.3 无先验知识

基于模型的离线强化学习方法需要学习显式的状态转移函数和奖励模型. 相比之下, 无模型方法无需知道先验状态转移函数和人类知识, 而是直接与环境交互来搜索同时满足成本约束和最大化回报的策略. 常用的方法为将安全约束融入到神经网络中实现从原始动作到安全空间的投影, 或者构建安全函数实现在安全区域内探索.

基于投影的方法. 此类方法旨在训练过程中确保约束满足. 当智能体生成的动作导致系统离开安全区域时, 则对该动作实施干预并将其投影到系统停留在安全区域内的最近动作. 为使强化学习算法在学习过程中不违反安全约束, Dalal 等[65] 直接在策略网络上增加一个安全层, 将原始网络输出的动作通过线性映射投影到一个安全集上, 从而在训练期间将智能体约束到受限区域中, 并实现零违反约束的目标. 然而, 该方法事先假设系统的安全性可通过在单个时间步长调整动作来保证, 因而无法确保全局安全性, 而且, 线性化近似方式可能无法很好地捕获复杂的环境动力学模型. 朱斐等[66] 在原始深度强化学习网络模型的基础上额外增加一个深度网络, 提出一种基于双深度网络的安全深度强化学习方法 (Dual deep network based secure deep reinforcement learning, DDN-SDRL). 其主要思路为: 首先, 依据安全性将状态划分为三种类型 (安全状态、临界状态、危险状态); 然后, 通过建立双经验池来分别存放危险样本 (导致任务失败的危险状态和临界状态样本) 和安全样本; 最后, 利用危险样本对新增的深度网络进行有针对性的训练并将训练结果作为惩罚项来改进原始深度网络 (由安全样本池为其提供训练样本) 的目标函数, 能够有效减少训练过程中智能体进入危险状态的次数, 从而在一定程度上增强了智能体的安全性. 但是, 通过对 6 个 Atari 2600 游戏上的仿真结果进行分析, 朱斐等[66] 指出: DDN-SDRL 存在波动范围大、训练不稳定的现象. 上述安全强化学习方法侧重于学习过程在原先策略网络的基础上添加投影步骤. 然而, 这类基于投影的方法需要在每个策略执行步骤中求解一个优化问题, 这就可能会导致高昂的计算成本, 并且在训练后去掉投影步骤时无法获得安全保证. 为此, Zheng 等[67] 提出了一种新的策略网络架构, 称为顶点网络 (Vertex network, VN), 通过将安全约束编码到策略网络架构中, 在探索和执行阶段都能保证安全. VN 中设计了一个新的安全层, 算法不解决投影优化问题, 而是在每个时间步骤计算安全区域的顶点, 并将动作设计为这些顶点的凸组合, 允许策略优化算法在训练期间只探索安全区域内部. 数值实验表明, 所提出的 VN 算法优于基于投影的强化学习方法. 约束成本函数类优化算法需要处理回报与成本之间的权衡, 同时易于陷入局部最优, 限制智能体的探索. 为解决这些问题, Marchesini 等[68] 并未将安全强化学习问题形式化为约束马尔科夫决策过程, 而是直接将进化算法融入到深度强化学习中, 提出安全导向搜索方法 (Safety-oriented search, SOS). 该方法定义了安全突变的概念, 利用已访问的不安全状态将探索偏向于更安全的动作, 并通过定义估计验证来刻画训练过程中的行为, 显著加速验证过程. 在 Safety Gym 基准数据集上进行测试, 结果表明 SOS 成功解决了回报与成本之间的权衡问题, 实现了与无约束算法类似的回报值, 并使得成本函数值最小, 获得了与约束算法类似的性能. 但是, 如果估计验证方法中属性选取不当, 则可能产生无法预料的行为.

构建安全函数. 此类方法利用先验知识建模从当前状态到安全状态的转换函数, 利用状态空间的结构将随机探索转变为安全探索. 通过引入安全控制和备份的概念, Mannucci 等[69] 提出了一种带有风险感知的安全探索算法 (Safety handling exploration with risk perception algorithm, SHERPA). 该算法不需要先验已知的全局安全函数以及动力学或环境模型, 而是依赖于智能体在探索阶段动力学的区间估计边界模型, 每次都以备份的形式生成一个临时安全函数, 将系统带到智能体过去已经访问过的状态附近, 并搜索满足接近性条件的策略动作, 从而提高安全性. 在简化的模拟四旋翼飞行器任务中显示了 SHERPA 可以有效避免危险状态. 但该算法仅适用于智能体知识有限的强化学习任务, 很难推广至复杂学习任务. 传统的 Q 学习通过 $\epsilon$-greedy 策略来实现随机探索, 其中决策者通过探索新的 Q 值来提高恢复决策的质量. 如果没有随机探索, Q 学习将无法保证正确收敛, 并且可能产生次优策略. 然而, 不受任何限制的随机探索可能导致算法选择不安全的动作, 这种破坏将对城市基础设施系统造成毁灭性打击. 为此, Memarzadeh 和

Pozzi[70] 提出了一种在无模型强化学习中引入基于模型的安全探索方法, 称为安全 Q 学习. 该方法对问题的状态空间结构进行建模, 并在 Q 学习基础上添加从当前状态到安全区域的动量函数, 从而改进动作选择策略, 使智能体实现安全探索. 在几个基础设施管理的例子中显示所提方法比传统的 Q 学习更接近最优性能, 缓解了随机探索所带来的风险. 然而与深度 Q 学习类似, 该方法仍存在训练不稳定和收敛性问题. 先前基于高斯过程的方法无法处理非平滑变化的安全问题, 为此, Wachi 等[71] 利用智能体观测得到的特征向量来预测安全函数值, 并提出了带有局部特征的安全策略优化算法 (Safe policy optimization with a local feature, SPO-LF), 能够在先验未知环境中优化安全策略. 该算法利用广义线性函数近似学习由传感器获得的局部可用特征与环境奖励/安全之间的关系, 同时优化智能体策略. 理论证明智能体能以很大的概率获得一个接近最优的策略, 同时在每个时间步上保证了安全约束. 实验结果进一步表明, 与现有的安全强化学习方法相比, SPO-LF 的样本复杂度和计算成本方面效率更高, 更适用于大规模问题. 然而, 该算法在具有连续状态和动作空间的实际应用上仍然有限.

### 2.2 修改学习目标

常规强化学习的目标是最大化长期回报, 忽略了危险状态对智能体造成的损害. 也就是说, 常规强化学习的目标函数中缺少对决策风险或损失的描述. 重要的是, 为了让强化学习智能体做正确的事, 目标函数必须准确地与想要实现的功能相匹配. 如果目标函数设计的不合理, 强化学习智能体就很有可能会面临安全问题. 由第 1 节可知, 安全强化学习问题可建模为约束马尔科夫决策过程, 然而, 由于回报目标函数和成本约束函数对智能体来说均是非凸的, 因此求解约束马尔科夫决策过程问题具有较大挑战性. 为求解约束最优化问题, 将修改学习目标类方法总结为两大类: 拉格朗日法和信赖域法. 总体思路均为引入风险相关信息, 将原先的约束优化问题转化为无约束最优化问题进行进一步求解.

#### 2.2.1 拉格朗日法

在求解约束最优化问题中, 拉格朗日法是最为常用的一种方法. 其基本思想为将约束马尔科夫决策过程问题转化为无约束对偶问题, 方法是对原始目标函数进行惩罚, 隐式表示安全约束, 通过拉格朗日乘子建立一个无约束的鞍点优化问题, 将约束问题转化为无约束问题, 然后交替地应用某种策略优化 (如策略梯度) 对其对偶变量进行更新. 因其思路简单, 易于实现, 在一些复杂任务中优于约束梯度方法, 拉格朗日法得以广泛研究.

Chow 等[72] 分别将风险表示为累积成本的机会约束和条件风险值, 构造了两种风险 CMDP 问题, 并利用拉格朗日乘子法将其转化为无约束优化问题加以求解. 该算法证明了其几乎必然 (依概率为 1) 收敛到局部鞍点, 但算法要求值函数和约束函数精确已知、可微且光滑, 该假设条件较为严格, 实际应用中难以满足.

以乘子网络为可行性指标, Ma 等[73] 提出了确保状态安全的可行演员−评论员 (Feasible actor-critic, FAC) 算法. 具体而言, 算法采用一个新增的神经网络 (乘子网络) 来近似关于状态的拉格朗日乘子, 实现从状态到乘子的映射, 通过状态互补松弛条件和乘子网络的梯度计算来准确指示状态的可行性. 利用原始−对偶梯度上升法训练策略和乘子网络, 得到一个确保每个可行状态安全的最佳可行策略、不可行状态的最安全策略、以及一个表明哪些状态不可行的乘子网络. 该算法能确保智能体通过优化策略渐近地达到安全行为, 但与其他安全强化学习算法一样, 该算法不能提供硬性保证, 而且与其他拉格朗日方法类似, FAC 仍存在训练不稳定以及对乘子网络更新敏感的问题.

传统的强化学习缺乏一种实用的方法来指定哪些行为是允许的或禁止的, 大多数情况下需要利用人类先验知识设定奖励函数来规范行为, 而 Roy 等[74] 旨在通过简单地指定阈值和指示函数来提供这些知识, 而不是要求专家示范或持续的人类反馈. 具体而言, Roy 等[74] 将无模型的软演员−评论员算法 (Soft actor-critic, SAC) 扩展为有约束的基于拉格朗日的 SAC 算法. 其中, 智能体的期望行为是由给定指标事件的发生频率来定义的, 将其视为 CMDP 中的约束. 该算法在 CMDP 框架中指定行为偏好, 在约束集中编码成功准则, 并使用拉格朗日方法自动权衡每个行为约束, 其中多约束问题的拉格朗日乘子通过 softmax 函数进行了归一化以提高训练期间的稳定性. 总体来说, 该算法研究了如何在同时遵守多个约束的情况下调整 CMDP 以解决基于目标的任务. 该方法可以看作是在不影响收敛要求的前提下, 在优化过程中放松行为约束的一种方法.

先前的安全强化学习方法大都考虑的是条件值风险约束, 针对平均约束, Sootla 等[75] 提出了一种使用状态增强的安全强化学习方法, 称之为 Sauté RL. 在具有平均约束的确定性环境中, 当某些初始状态的安全成本较高而同时满足平均约束的其他初始状态的成本较低时, 可能会导致不必要的影响.

所提方法通过确保所有初始状态都满足相同的约束来避免这种情况. 该方法将安全约束条件纳入到状态空间中, 对每一个受控轨迹都强制执行该约束条件, 因此可以依概率为 1 地满足安全约束. 此外, 该方法满足贝尔曼方程, 并更接近于解决几乎必然满足约束的安全强化学习问题. 所提方法具有即插即用的性质, 可以与现有的很多无模型算法或基于模型的方法相结合. 然而, Sauté RL 通过增加约束数量来增强状态空间, 而理论采样效率取决于状态空间的维度, 因此无法适用于高维控制问题. 而且 Sauté RL 仍然没有解决训练过程中的约束违反问题, 其成本函数中仍隐含着可能违反安全约束的信息.

Tessler 等[76] 利用拉格朗日乘子法将约束作为惩罚信号引入回报函数, 提出一种约束型演员−评论员 (Actor-critic) 方法, 称为回报约束策略优化 (Reward constrained policy optimization, RCPO). RCPO 采用了多个时间尺度: 在快时间尺度上对 critic 进行更新, 使用时序差分学习估计带惩罚的回报函数; 在中间时间尺度上对 actor 进行更新, 使用策略梯度方法学习策略; 在慢时间尺度上对拉格朗日乘子进行更新, 通过缓慢增大惩罚系数来满足约束. RCPO 为原始−对偶方法提供了渐近收敛分析, 建立了局部收敛保证, 确保收敛到不动点, 但多时间尺度方法涉及到的多个学习率在实际使用中通常难以调整.

对于采用函数近似的大规模强化学习问题而言, 约束马尔科夫决策过程的目标函数和约束条件均被建模为关于策略参数的非凸函数. 为求解非凸约束优化问题, Yu 等[77] 直接将非凸函数近似为从策略梯度估计器获得的凸二次函数, 从而转化为求解一系列凸约束优化子问题, 提出一种具有收敛保证的约束策略梯度算法. Yu 等[77] 证明了这些子问题得到的策略参数几乎必然收敛到原始非凸问题的驻点. 同时, 约束强化学习具有零对偶间隙性, 这为对偶域中的策略梯度算法提供了理论保障.

违反约束条件在实践中可能会带来灾难性的后果, 如何在不违反约束的情况下实现最优目标是一个重要挑战. 为此, Bai 等[78] 设计了一种保守的随机原始−对偶算法 (Conservative stochastic primal-dual algorithm, CSPDA) 来解决 CMDP 问题 (等价于鞍点问题), 并利用遗憾分析证明了约束违反为零.

拉格朗日法将约束优化问题简化为带有辅助惩罚项的无约束优化问题, 简单的拉格朗日方法就能找到满足约束的策略, 而且获得较大的回报. 尽管当策略渐近收敛时拉格朗日法能够确保部署安全性, 但仍存在以下弊端: 一个鞍点优化问题就相当于一系列马尔科夫决策过程求解问题, 计算量较大; 对拉格朗日乘子的初始值和学习率比较敏感, 因此超参数调优过程会产生很大开销; 拉格朗日乘子和策略参数都不为零时问题的目标非凸非凹, 迭代求解可能无法保证解的收敛速度; 在训练过程中无法保证生成策略的安全性.

#### 2.2.2 信赖域法

与拉格朗日法不同的是, 信赖域法显式表示安全约束, 通过修改信赖域策略梯度来求解约束策略优化问题, 在每次迭代过程中将策略投影到一个安全的可行集内, 从而确保策略在预期约束范围内.

在对 CMDP 优化求解的过程中, 需要考虑长期而非单步成本函数满足约束条件. 为此, Achiam 等[79] 对信赖域策略优化 (Trust region policy optimization, TRPO)[80] 方法进行了拓展, 率先提出一种用于安全强化学习的通用策略搜索方法, 称为约束型策略优化 (Constrained policy optimization, CPO), 可以确保智能体在学习过程中的每一步都满足约束条件. CPO 的主要思路为: 首先, 鉴于原始 CMDP 优化问题的目标函数非凸、不连续, 构造一个代理函数来近似原始目标函数; 其次, 把目标函数和成本函数通过泰勒二阶展开, 得到一个简化的目标函数; 然后, 简化目标函数为凸优化问题, 可通过对偶问题或者直接根据最优性条件得到策略的迭代表达式; 最后, 鉴于近似误差可能导致求得的解违反约束, 利用回溯线搜索以保证满足约束. 但 CPO 采用二阶泰勒展开式逼近目标函数和成本函数的做法会导致 Fisher 信息矩阵的规模较大, 因此需要考虑采用共轭梯度法来间接计算 Fisher 信息矩阵的逆. 而且, 违反约束后利用回溯线搜索的更新规则可能会减缓策略学习的进度.

借助投影梯度下降法的思想, Yang 等[81] 提出了一种基于投影的约束策略优化 (Projection-based constrained policy optimization, PCPO) 方法. 该方法分两个阶段来更新策略: 首先执行无约束更新, 采用 TRPO 方法优化回报函数, 得到中间策略, 然后利用 KL (Kullback-Leibler) 散度将策略投射回约束集上来调节违反约束的情况, 即在约束集中选择最接近中间策略的可行策略. 与 CPO 相比, PCPO 首先优化回报函数, 然后使用对约束集的投影来保证整个学习过程中满足约束, 因此可以在保证安全的同时对回报函数进行优化. 但 CPO 和 PCPO 存在的共性问题都是原始问题的解析解需要涉及 Fisher 信息矩阵求逆, 对于大型约束马尔科夫决策过程而言, 矩阵求逆计算量大, 近似方法又

会产生计算误差.

为减小矩阵逆近似计算带来的误差, Zhang 等[82] 提出一种一阶约束优化方法 (First order constrained optimization in policy space, FOCOPS). 主要思路为: 使用原始–对偶梯度方法来求解带有成本约束的信赖域问题, 首先在非参数化策略空间中解决一个带约束的优化问题, 然后再将更新策略映射回参数化策略空间. 相较于 CPO 和 PCPO, 所提方法只采用了线性近似, 无需求解 Fisher 逆矩阵, 因此计算效率较高. 高维连续控制任务上的仿真结果表明在近似满足约束的情况下, 使用一阶近似性能上优于复杂的二阶近似方法 (如 CPO), 但并未从理论层面对这一观测结果进行论证.

同样考虑到逆矩阵难计算的问题, Zhang 等[83] 将传统的近端策略优化算法拓展为安全强化学习方法, 提出惩罚近端策略优化 (Penalized proximal policy optimization, P3O) 算法. 该算法利用精确惩罚函数将成本约束转化为无约束优化问题并采用一阶优化进行求解, 将 PPO[84] 中的裁剪代理目标扩展到 CMDP 中来消除信赖域约束. 所提算法避免了二次近似和高维 Hessian 矩阵求逆, 有利于求解使用深度神经网络的大型 CMDP 问题.

原始–对偶问题求解 CMDP 会引入额外的对偶变量, 为此, Xu 等[85] 提出了第一个对全局最优具有可证明收敛性保证的基于原问题求解的安全强化学习方法, 即约束修正策略优化 (Constraint-rectified policy optimization, CRPO) 算法. 其中所有的策略更新均采用自然策略梯度, 并在原始域中进行. 若不违反约束, CRPO 通过求解无约束最大化奖励目标来更新策略, 否则, 通过无约束最小化约束目标函数项进行更新策略, 沿着被违反约束的下降方向, 将策略瞬时修正回约束集. 由于基于原问题求解的方法无需引入额外的对偶变量来优化, 因此涉及的超参数调优较少, 但算法收敛速度较慢, 且在训练阶段无法保证安全性.

尽管原始–对偶方法已广泛用于求解约束优化框架, 但其存在训练不稳定和缺乏最优性保证等问题. 为此, Liu 等[86] 从概率推理的角度出发, 提出了约束变分策略优化 (Constrained variational policy optimization, CVPO) 算法. CVPO 采用期望最大化算法进行求解, 将安全强化学习问题分解为两个阶段: 1) 凸优化学习阶段, 该阶段采用非参数变分分布, 并具有最优性保证; 2) 有监督学习阶段, 该阶段采用信赖域正则化策略改进方法, 并具有稳定性保证. 在连续控制任务上的实验表明, 与基于原始–对偶的安全强化学习方法相比, CVPO 训练更稳定、采样效率更高. 但文献 [86] 中约束阈值的选取以及选取准则并不明确, 且涉及的超参数多, 难以用于实际场景.

上述方法均为用于约束型强化学习的信赖域方法, 在每一次策略更新中都近似地强制执行约束条件, 因此策略在训练过程中能确保安全性. 由于这些方法与 TRPO 紧密相关, 将其应用于 PPO 类方法进行约束优化很简单, 但尚不清楚如何将其与不属于近端策略梯度类型的方法结合使用, 例如深度确定性策略梯度算法. 除此之外, 此类方法还存在一些不足之处: 非凸策略优化的凸近似会产生不可忽略的近似误差, 因而采用一阶或二阶近似的方法只能学到接近满足约束的策略; 当原问题在一定初始策略下不可行时, 需要采用额外的恢复方法, 通过与环境交互将策略恢复到可行集中, 因此采样效率低; 二阶近似涉及矩阵求逆, 高维环境下计算代价大, 无法适用于求解大规模约束马尔科夫决策过程问题.

### 2.3 离线强化学习

上述两大类算法 (修改学习过程和修改学习目标) 可归结为在线强化学习, 即智能体需要与环境进行不断交互来学习如何执行任务. 离线强化学习要求智能体完全从静态的离线数据集中学习而不进行探索, 因此从数据层面确保了智能体的训练安全性[19−20]. 但在策略部署阶段, 此类方法并未考虑任何风险相关因素, 因而无法确保部署时的安全性. 监督学习假定训练数据与测试数据独立同分布, 从而在训练集上学到的模型能够在测试集中拥有很好的性能. 然而, 离线强化学习通常面临所学习的策略 (目标策略) 与从离线数据集中观测到的策略 (行为策略) 分布不同而造成分布偏移问题[19]. 针对该问题, 学者们从策略约束、值约束和预训练模型等多角度进行研究, 提出了一系列离线强化学习方法.

#### 2.3.1 策略约束

策略约束类方法通常利用生成模型 (如变分自编码器、生成对抗网络) 在隐空间对动作进行限制, 或者利用分布度量 (如 KL 散度、最大均值差异) 将目标策略限制在行为策略的分布范围内, 目的都是使目标策略近似行为策略分布, 从而缓解分布偏移问题.

Fujimoto 等[87] 提出首个离线强化学习算法, 称为批约束深度 Q 学习 (Batch-constrained deep Q-learning, BCQ), 它能从任意批数据中学习而无需探索. BCQ 限制了动作空间, 利用变分自编码器模型来生成与离线数据集分布相近的动作, 确保只选

择类似于行为策略选择的动作来约束目标策略. 同时, 结合一个扰动模型对生成的动作进行调优, 使动作具有多样性. 实验表明, 相比于模仿学习, BCQ 在连续控制任务中具有独特优势和巨大潜力. 然而, 严格限制目标策略接近行为策略未必在所有场合下都奏效. 当离线样本数不足时, BCQ 将会受到行为策略分布密度的限制, 无法很好地对先前从未见过的分布外轨迹进行拟合.

Kumar 等[88] 则认为自举误差是当前方法不稳定的一个关键来源, 其中自举误差是指训练过程中分布外动作通过贝尔曼回溯进行累积而引起的误差, 并通过理论分析了在 Q 学习过程中约束动作可以减少误差传播. 同时, 为减少分布外动作的影响, Kumar 等[88] 引入了支撑集的概念, 并提出一种隐式策略约束的离线强化学习算法 BEAR (Bootstrapping error accumulation reduction). BEAR 的目标并不是显式约束所学策略与行为策略越像越好, 而是将所学策略保持在行为策略的支撑集范围内, 约束两个策略分布之间的最大均值差异距离小于某个阈值, 这样就能通过学习策略的上界可集中性来控制误差传播, 同时减小与最优策略分布之间的距离. 与 BCQ 相比, BEAR 在限制所学策略方面没有太过保守. 在一系列连续控制任务上的仿真结果表明, BCQ 仅在由专家策略收集而来的离线数据集上表现良好, 而 BEAR 在次优策略甚至是随机策略数据集上会优于 BCQ. 但 BCQ 和 BEAR 这类策略约束方法在很大程度上仍依赖离线数据集的质量, 当数据质量不高时, 此类方法总体仍无法获得满意的结果.

BCQ 使用生成模型近似数据集分布, 并在生成的动作集合中手动选择使 Q 值最大的动作. PLAS (Policy in the latent action space) 方法[89] 是对 BCQ 的进一步扩展, 该方法不是在动作空间上学习扰动模型, 而是通过极大似然学习生成模型并在潜在空间中学习确定性策略. 假设潜在动作空间隐式定义了对动作输出的约束, 从而训练时策略是在数据集支持范围内选择动作, 而不受数据集分布密度的限制. 在各种连续控制任务中的实验验证了所提方法的性能. 然而, Chen 等[90] 指出 PLAS 只能模拟数据集中低回报但高密度的样本, 而无法捕获高回报动作.

为此, 针对异构离线强化学习任务, Chen 等[90] 提出潜在变量优势加权的策略优化 (Latent-variable advantage-weighted policy optimization, LAPO) 算法. 该算法在训练策略与最大化数据的优势加权对数似然之间交替学习潜在空间策略并实现奖励最大化, 利用潜在变量生成模型来表示高优势的状态–动作对, 使得强化学习策略偏向于选择训练数据支持的动作, 同时有助于解决任务的数据分布偏移问题. LAPO 在结构上与 PLAS 类似, 因为这两种算法都需要学习生成模型和潜在空间策略. 然而, BCQ 和 PLAS 中的生成模型是通过预先训练来近似整个数据集上的分布, 并且在训练潜在策略时是固定的, 这可能会限制模型在数据集中样本数较少时对高回报样本的表达能力. 与先前工作不同的是, LAPO 通过交替学习生成模型、优势函数和潜在策略来训练优势加权生成模型, 从而能够在潜在空间中利用简单的高斯分布捕获高回报动作. 仿真结果表明, LAPO 在异构数据集上明显优于 PLAS.

#### 2.3.2 值约束

值约束常用的方法如下: 在值函数基础上添加正则化项, 使其估计值更加保守; 利用估计的不确定性来学习一个悲观值函数, 从而避免产生分布外动作; 通过重要性采样, 用行为策略的样本来评估目标策略. 其目的都是基于悲观/保守的思想, 通过限制值函数, 从而缓解分布偏移问题.

目标策略与行为策略之间的分布偏移问题导致传统的异策略强化学习方法出现值函数过估计. 为解决此问题, Kumar 等[91] 提出保守的 Q 学习 (Conservative Q-learning, CQL) 算法. 该算法通过向其目标函数添加值函数正则化项来学习真实 Q 函数的下界, 使得该策略下的 Q 函数期望值低于其真实值, 从而防止由于分布外动作和函数近似误差而导致的高估问题. 在离散和连续控制域上的实验结果表明, CQL 的性能大大优于现有的离线强化学习算法, 尤其适宜处理复杂的多模态数据分布. 与监督学习方法类似, 离线强化学习也极易出现过拟合现象, 因此设计简单有效的提前终止方法是未来的一项重要挑战.

在离线环境下满足安全约束并非易事, 这是由于所学策略与离线数据集之间可能存在分布偏移, 导致安全约束值错误估计, 因此只能学到次优解. 为此, Xu 等[92] 提出了一种安全离线强化学习算法, 称为约束惩罚的 Q 学习 (Constraints penalized Q-learning, CPQ). 该算法不使用显式的策略约束, 也不会受到数据集分布密度限制, 允许使用混合行为策略产生的数据集, 并通过修改奖励 critic 的贝尔曼更新方程来惩罚不安全的状态动作对. 连续控制任务的仿真结果表明, CPQ 能够在满足安全约束的同时获得最大回报. 总而言之, CPQ 在离线情况下训练强化学习算法, 为现实任务中的安全与高质

量控制提供了可靠的策略保证.

目前大多数离线强化学习方法在训练过程中需要评估未见过的动作值来改进策略, 因此需要将这些动作限制在分布之内, 或者对它们的值进行正则化. Kostrikov 等[93] 提出隐式 Q 学习 (Implicit Q-learning, IQL), 在训练过程中完全避免评估数据集以外的动作, 同时仍然能够进行多步动态编程. 该算法主要贡献在于在策略评估步骤中, IQL 没有使用从行为策略中采样的目标动作来更新 Q 函数, 而是针对每个状态的数据集动作分布来近似值分布的上期望值. 通过交替使用期望回归拟合此值函数, 然后用它计算贝尔曼备份来训练 Q 函数. 该算法实现简单且计算效率高.

基于重要性采样的方法直接估计状态边际重要性比率 (密度比), 得到无偏值估计, Zhang 等[94] 提出广义平稳分布校正估计 (Generalized stationary distribution correction estimation, GenDICE) 算法. 该算法将约束条件扩展为状态–动作的边际重要性比率, 并直接优化其修正的贝尔曼方程所对应的剩余误差. GenDICE 能同时处理给定多个行为不确定样本的折扣平稳分布和平均平稳分布问题. 然而, 该算法所涉及的优化问题不是凸凹鞍点问题, 因此无法确保收敛或找到期望解.

确保离线强化学习的采样效率通常需要满足两个强假设条件. 1) 数据覆盖 (所有策略的集中性): 离线数据分布 (在技术意义上) 对所有候选策略诱导的状态分布提供了良好覆盖; 2) 函数近似 (贝尔曼完备性): 值函数类在贝尔曼最优性算子下闭合. Zhan 等[95] 放松了这两个假设条件并提出一种原始–对偶正则化的离线强化学习 (Primal-dual regularized offline reinforcement learning, PRO-RL) 算法, 其中对偶变量 (目标策略的折扣占有率) 使用边际重要性比率对离线数据进行建模. 该算法在不满足贝尔曼完备性的情况下, 利用正则化线性规划求解异策略评估. 但由于引入正则化, PRO-RL 所学到的策略一般是次优的.

#### 2.3.3 预训练模型

借助模仿学习或元学习的思想, 对离线数据进行预训练来初始化模型, 使之获得一个较好的初始策略, 从而快速适应新环境.

BCQ 和 BEAR 这类算法都是利用生成模型对行为策略进行建模, 若离线数据集由随机策略或噪声策略生成时, 则很难对行为策略进行精准预测. 为此, Siegel 等[96] 利用先前学过的先验知识, 提出优势加权行为模型 (Advantage-weighted behavior model, ABM). 具体而言, 采用策略迭代方法, 通过学习先验知识了解哪些候选策略可能在离线数据集支持范围内, 从而形成行为数据的优势加权模型; 再强制执行策略提升步骤, 使所学策略倾向于先前执行过的经验丰富的动作, 这些动作可能在新任务上取得更好的性能. ABM 可以使用任意行为策略生成的数据, 能够在冲突的数据源中稳定学习, 在连续控制基准与机器人多任务学习的测试中验证了所提算法优于现有的两个离线强化学习算法 BCQ 和 BEAR.

大多数模仿学习方法关注的是过滤掉次优行为, 随后再利用传统的监督回归损失进行处理. Wang 等[97] 提出 critic 正则化回归 (Critic regularized regression, CRR) 方法, 从离线数据中学习策略. 该方法使用指示函数过滤掉低于平均水平的动作, 而选择了一个更悲观的优势估计. CRR 的策略更新可看作是行为克隆的加权版本, 其中权重由 critic 网络所确定. critic, 即状态–动作值函数, 则是通过对奖励标注的数据进行 Q 学习训练而来的. 仿真结果表明, 对于具有高维状态和动作空间的任务, CRR 能表现出更好的性能.

离线强化学习无需与环境进行实时交互, 仅在已收集的离线轨迹数据集上进行训练, 这与监督学习的方式极为相似. 基于该思想, Emmons 等[98] 直接将策略学习问题视为监督学习问题, 并提出 RvS (Reinforcement learning via supervised learning) 算法. 该算法使用条件行为克隆的方式来学习策略, 其中样本由观测到的蒙特卡洛回报进行加权、过滤或条件化, 从而避免了对值函数进行估计. 仿真结果表明, 纯监督学习 (最大化数据中观察到的动作的可能性) 方式表现的与保守时序差分学习方法一样好; 在广泛的任务下, 简单的前馈模型可以与先前序列模型的性能相匹配; 若条件变量 (如目标或者奖励值) 选择恰当, 则 RvS 可以获得最优性能. 然而, 条件变量的选取对算法性能产生很大影响, 如何自适应选择条件变量以提高 RvS 算法的适用性是未来值得研究的方向.

Uchendu 等[99] 使用离线数据、示范或预先存在的策略来初始化强化学习策略, 提出一种快速启动的强化学习 (Jump-start reinforcement learning, JSRL) 方法. 该方法采用了两种策略: 指导策略和探索策略. 智能体在学习过程开始时使用指导策略而不是随机策略作为探索策略的初始起点, 随着探索策略的不断更新, 指导策略的作用逐渐减弱, 因此加快了学习过程, 也提高了算法的泛化性能. 视觉机器人实验表明 JSRL 明显优于先前提出的模仿和强化学习方法.

除了改善单个强化学习任务的性能外, 学者们

也逐步转向对模型的通用性和泛化性能进行研究, 力求找到一种实用的强化学习范式, 能够从离线数据中学习以快速适应新任务, 这便是离线元强化学习. 在离线元强化学习的设定中, 引入了任务分布学习的概念, 离线数据集不再是仅从单个任务中收集而来, 而是通过不同的行为策略从多个任务中获取得到. 训练智能体的目的是学习一种元策略, 能够有效地适应从未见过的任务. Zhao 等[37] 提出带有示范自适应的离线元强化学习 (Offline meta-RL with demonstration adaptation, ODA) 算法. 该算法首先使用离线数据进行预训练, 从中学习出一种自适应策略, 再使用基于少量用户提供的示范数据来快速适应新任务, 最后通过在线微调做进一步的适应. 仿真结果表明, 所提算法仅需少量训练样本即可快速适应各种不同的工业连接器插入任务.

尽管上述方法在一定程度上解决了分布偏移问题, 其得以成功应用的前提假设是离线数据可以充分覆盖高奖励的状态动作对. 在现实世界场景中, 数据收集可能是昂贵的或受限的, 实际应用中离线数据集的状态动作空间覆盖可能相当狭窄, 这也是离线强化学习当前面临的一大挑战. 由于仅从离线数据中训练, 因此上述所有的离线强化学习算法都能做到训练时安全. 只有 CPQ[92] 考虑了安全约束, 因此能实现部署时安全, 而其余的传统离线强化学习算法在不考虑安全性的情况下无法确保部署时安全.

## 3 基准测试平台

安全强化学习和基于安全学习的控制方法已得到广泛的研究, 为了充分评估和公平比较现有的安全强化学习算法, 并促进其在现实业务场景中的落地, 学者们开发了多种基准测试平台. 本节介绍 5 个开源工具, 这些工具包含了具有安全要求的模拟环境, 并集成了常用的安全强化学习算法, 对定量评估现有安全强化学习算法具有积极的指导意义. 5 种基准测试平台的任务如图 1 所示, 适用条件与特性如表 2 所示.

### 3.1 Safety Gym

OpenAI 团队最先推出一套加速安全探索研究的工具 Safety Gym[100]. 该工具包基于 MuJoCo 物理引擎, 采用 OpenAI 的 Gym 接口, 同时与许多当前的强化学习库无缝集成. Safety Gym 由两部分组成: 1) 一个环境创建器, 允许用户通过混合和匹配各种物理元素、目标和安全要求来创建一个新的环境; 2) 一套预先配置的基准环境, 用于帮助标准化定量评估安全强化学习方法的好坏. Safety Gym 开源代码库: https://github.com/openai/safety-gym.

#### 3.1.1 环境设置

在所有 Safety Gym 环境中, 机器人必须在复杂环境中进行导航才能完成任务. 有三个预先设定的机器人 (Point、Car、Doggo), 三个主要任务 (Goal、Button、Push), 每个任务都有两个难度级别.

#### 3.1.2 基准算法

为便于科研人员快速上手, Safety Gym 还提供了一些基准安全强化学习算法, 例如: PPO[84]、TRPO[80]、PPO 和 TRPO 的拉格朗日惩罚版本[101]、CPO[79] 等. 基准算法开源代码库: https://github.com/openai/safety-starter-agents.

### 3.2 safe-control-gym

多伦多大学航空航天研究所团队提出了一个开源基准工具 safe-control-gym[102], 用于评估基于学习的控制和安全强化学习算法性能. 该模拟环境基于 Bullet 物理引擎, 采用 OpenAI 的 Gym 接口, 同时与许多当前的强化学习库无缝集成. safe-control-

表 2 安全强化学习基准测试平台对比

Table 2 Comparison of benchmarking platforms for safe reinforcement learning

| 基准测试平台 | 任务类型 | 适用方法 | 基准算法类型 | | 特点 |
|---|---|---|---|---|---|
| Safety Gym | 机器人导航 | 修改学习过程与目标 | 无模型方法 | 同策略 | 包含多个高维连续控制任务, 使用最广泛的安全强化学习算法评估平台 |
| safe-control-gym | 机器人控制 | 修改学习过程与目标 | 无模型方法与基于模型的方法 | 同策略与异策略 | 能实现基于模型的方法, 可以方便地与控制类方法进行对比 |
| SafeRL-Kit | 自动驾驶 | 修改学习过程与目标 | 无模型方法 | 异策略 | 首个针对自动驾驶任务的异策略安全强化学习算法基准测试平台 |
| D4RL | 机器人导航与控制、自动驾驶 | 离线强化学习 | 无模型方法 | 离线学习 | 收集有多个环境的离线数据, 已成为离线强化学习算法的标准评估平台 |
| NeoRL | 机器人控制、工业控制、股票交易、产品促销 | 离线强化学习 | 无模型方法与基于模型的方法 | 离线学习 | 包含多个高维或具有高度随机性的现实应用场景任务 |

gym 能够支持基于模型的方法、表达安全约束以及捕捉现实世界中的非理想状态 (如不确定的物理属性和不完美的状态估计), 对应于三个核心特征: 先验符号模型、约束规范和干扰注入, 以便同强化学习与控制方法无缝衔接. safe-control-gym 开源代码库: https://github.com/utiasDSL/safe-control-gym.

#### 3.2.1 环境设置

safe-control-gym 包括三个动力系统: 推车杆 (Cart-pole), 一维和二维四旋翼飞行器 (Quadrotor), 两种控制任务: 稳定和轨迹跟踪.

#### 3.2.2 基准算法

safe-control-gym 包括:

1) 控制与安全控制基准算法. 对于标准状态反馈控制器, 包含线性二次调节器 (Linear quadratic regulator, LQR) 和迭代 LQR (Iterative LQR, iLQR) 两种方法[103]. 对于预测控制基准, 包含线性模型预测控制 (Linear model predictive control, LMPC) 和非线性模型预测控制 (Nonlinear model predictive control, NMPC) 两种算法[104].

2) 强化学习基准算法. 两个无模型的方法 PPO[84] 和 SAC[105].

3) 基于安全学习的控制基准算法. 提供一种基于高斯过程的模型预测控制 (Gaussian process-based model predictive control, GP-MPC) 方法[106], 该方法采用高斯过程对不确定的动态进行建模, 用来更好地预测系统的未来演变, 以及根据预测范围内动力学的置信度来收紧约束.

4) 安全与鲁棒的强化学习基准算法. 对于安全强化学习, 将基于安全层的方法[65] 添加进 PPO[84] 中用于测试. 对于鲁棒强化学习, 包含两种基于对抗学习的方法 RARL[107] 和 RAP[108].

5) 学习控制器的安全认证基准算法. 一个常用的安全过滤器方法, 即模型预测安全认证 (Model predictive safety certification, MPSC)[109], 以及控制障碍函数[110].

### 3.3 SafeRL-Kit

最近, Zhang 等[17] 提供了面向自动驾驶领域的安全强化学习基准测试工具包 SafeRL-Kit. 在两个自动驾驶环境中进行实验, 所有基准算法均采用统一框架, 方便进行对比. SafeRL-Kit 开源代码库: https://github.com/zlr20/saferl_kit.

#### 3.3.1 环境设置

SafeRL-Kit 在两个模拟环境中进行基准测试, 包括一个四驱赛车速度控制环境 SpeedLimit[111] 和一个轻量化的真实汽车自动驾驶仿真平台 MetaDrive[112].

#### 3.3.2 基准算法

该工具包实现了几种最新的安全强化学习算法, 包括用于安全校正的安全层方法[65]、用于安全恢复的恢复强化学习 (Recovery RL) 方法[18]、异策略拉格朗日 (Off-policy Lagrangian) 方法[113]、FAC[73] 以及 Zhang 等[17] 提出的用于安全约束的精确惩罚优化 (Exact penalty optimization, EPO) 方法. 所有方法均采用统一的异策略 actor-critic 网络架构进行实现, 从而提高了采样效率, 且方便与人类示范先验知识相结合.

### 3.4 D4RL

针对离线强化学习, Fu 等[114] 提出了一个深度数据驱动的强化学习开源基准数据集 D4RL. 该工具包基于 MuJoCo 物理引擎, 采用 OpenAI 的 Gym 接口. D4RL 提供了来自不同领域的多项任务, 涵盖导航、机器人操作和自动驾驶等应用领域, 包括几种基准离线强化学习算法, 并提供了非常简单的 API 接口, 方便学习者直接获取离线数据集来实现算法的训练. D4RL 开源代码库: https://github.com/rail-berkeley/d4rl.

#### 3.4.1 环境设置

D4RL 包含了来自 7 个不同领域的 42 项任务.

1) 导航任务, 包括 Maze2D 和 AntMaze 分别提供的 3 种和 6 种迷宫导航任务, 以及 CARLA 模拟器提供的 2 种基于现实世界视觉的导航任务.

2) Gym-MuJoCo 作为 OpenAI 基准测试的 HalfCheetah、Hopper 和 Walker2D 三个数据集的共计 12 项任务.

3) 机器人操作任务, 包括 Adroit 和 Franka-Kitchen 平台分别提供的 12 项和 3 项任务.

4) 自动驾驶任务, 包括 Flow 基准测试的 4 项任务.

#### 3.4.2 基准算法

该工具包涵盖了多种离线强化学习算法, 例如: 策略约束类方法 BCQ[87]、BEAR[88]、BRAC-p[115]、BRAC-v[115] 和 AWR[116], 值函数正则化类方法 CQL[91] 和边际重要性采样方法 AlgaeDICE[117].

### 3.5 NeoRL

最近, 国内南栖仙策团队提出了离线强化学习基准数据集 NeoRL[118], 其中数据集是由更保守的策略收集而来. 除了广泛使用的运动控制任务外, NeoRL 还包含了一些高维或具有高度随机性的现

实场景任务, 如工业控制、金融交易等. NeoRL 开源代码库: http://polixir.ai/research/neorl.

#### 3.5.1 环境设置

该基准数据集由 5 大领域的 52 项任务构成.

1) Gym-MuJoco 环境中包含 3 类连续控制任务: HalfCheetah、Hopper、Walker2d, 由保守的策略生成且数据量有限.

2) IB 为工业基准领域, 用于模拟各种工业控制任务中呈现的特性, 如风力或燃气轮机、化学反应器等.

3) FinRL 模拟股票交易市场, 包含股票池中的 30 只股票和过去 10 年的交易历史.

4) CityLearn 用于控制不同类型建筑的储能来重塑电力需求的聚集曲线. IB、FinRL 和 CityLearn 这三个领域数据集都来自高维连续状态与动作空间, 且具有高随机性.

5) SalesPromotion 模拟真实商品促销平台, 平台运营商的目标是使总收入最大化, 该领域的数据集由人工操作员和真实用户提供.

#### 3.5.2 基准算法

包括行为克隆 BC, 无模型离线强化学习算法 BCQ[87]、PLAS[89]、CQL[91] 和 CRR[97], 以及基于模型的离线强化学习算法 BREMEN[119] 和 MOPO[120].

## 4 应用领域

安全强化学习在诸多领域都有广泛的应用前景, 也产生了许多成功的案例. 下面从自动驾驶、机器人控制、工业过程控制、电力系统优化和医疗健康 5 大领域展开详细叙述.

### 4.1 自动驾驶

自动驾驶是人工智能的一个发展产物. 驾驶自动汽车时不仅要求高效准确到达目的地, 还应做到安全可靠. 因此, 智能体如何在性能与风险之间权衡是当前该领域面临的最大挑战. 针对自动驾驶的安全性问题, 许多新颖的算法被提出, 也取得了一定成就.

针对无人车的安全探索问题, 代珊珊和刘全[30] 提出了一种基于动作约束的 SAC 算法. 通过在回报函数中引入惩罚项来避免无人车陷入危险状态, 并对无人车的动作进行限制, 使之避免发生碰撞或偏离轨道. 与传统的 SAC 算法相比, 所提算法在自动驾驶车道保持任务中可以有效规避风险.

为实现自动驾驶车辆在典型的多变交互场景 (即环岛) 下的自适应决策, Zhang 等[23] 提出优化嵌入强化学习算法. 所提算法对 actor-critic 框架中的 actor 网络进行改进, 新增了任务表示的状态向量来重组策略网络, 以便智能体能适应不同类型的环岛. 环岛场景下的仿真结果表明, 所提算法可以通过在线调整期望加速度和动作时间来实现自适应决策, 避免交互场景中的突发事件, 确保安全行驶.

自动驾驶车辆在正常行驶过程中经常会遇到上下匝道、车道合并、或由于道路施工等因素引发的车辆汇流, 如何有效提升该场景下车辆通行的效率及安全性, 是自动驾驶决策系统开发中的一个关键问题. 常用的方式为通过车辆之间相互协同通信来解决道路冲突问题. 为避免在汇流场景下自动驾驶车辆的碰撞行为, Kamran 等[12] 结合强化学习与基于优化的轨迹规划器, 将汇流场景下车间协同问题建模为部分可观测马尔科夫决策过程. 为了获得最优的行为策略, Kamran 等[12] 通过引入可扩展的深度 Q 网络来预测其他驾驶员的意图并选择最佳动作. 仿真结果表明, 所提算法可以从车辆的历史状态中有效地识别出协同车辆, 并产生交互式的操作, 从而使自动驾驶更为舒适. 同时, 由于规划器内部含有安全约束, 因此所提出的算法可以确保车辆安全无碰撞.

为解决车辆与行人交互场景下的安全问题, 即避免车辆向前行驶时与试图过马路的行人相撞, Trumpp 等[13] 为安全自主车辆开发了新的行人防撞 (Pedestrian crash avoidance mitigation, PCAM) 系统, 并提出了一种基于深度多智能体强化学习的人行横道车辆–行人交互建模方法. 同时, 根据智能体的碰撞率和由此产生的交通流效率对所开发的 PCAM 系统进行评估, 以便观察行人行为的不确定性或噪声对智能体决策的影响. 仿真结果表明, 所提方法能有效缓解碰撞问题, 自主车辆学会了更智能的过马路行为.

### 4.2 机器人控制

机器人控制可以渗透到生活的各个场景, 例如, 训练机器人行走、导航、路径规划等辅助人类完成复杂任务. 在工业领域, 在确保安全的前提下利用机器人控制使效益最大化. 传统强化学习方法需要实体机器人通过不断试错的方式进行数据收集, 这种做法昂贵且耗时, 确保机器人的安全性是一项迫在眉睫的挑战[121]. 下面总结了近些年安全强化学习在机器人控制方面的应用.

当训练两足机器人行走时, 尤其是完成具有一定强度的任务 (比如快速行走) 时, 由于其自身不稳定而导致在学习过程中很容易摔倒. 为此, García 和 Shafie[27] 提出基于策略重用的安全强化学习算

法, 来改善人形机器人的行走行为. 该算法假设存在一个安全的基线策略, 允许人形机器人行走, 并以概率方式重用这种策略来学习一个更好的策略, 其中遵循基线策略的概率由一个单调递增风险函数来确定. 在真实的人形机器人上的仿真结果显示所提算法能极大提高学习速度, 同时减少学习期间的摔倒次数.

同样针对机器人行走安全问题, 为降低四足机器人行走过程中腿部运动的损害风险, Yang 等[14] 提出基于双重策略的安全强化学习框架. 该框架包含两种策略: 安全恢复策略, 其将机器人从接近不安全的状态中恢复过来; 学习者策略, 其为执行预期控制任务而进行优化. 算法在这两种策略之间进行切换, 从而防止智能体违反安全约束, 同时尽量减少对学习过程的干预. 在模拟真实的四足机器人实验中验证了所提框架能够有效减少机器人摔倒次数, 提高了安全性和学习效率.

为解决机器人在无地图导航任务下的安全性问题, Corsi 等[35] 将基于场景的编程与特定于机器人环境中的约束强化学习系统相结合. 在 Lagrangian-PPO 算法[100] 的训练过程中融入领域专家知识, 从而确保智能体产生既安全又高性能的可信赖策略. 在机器人无地图导航仿真平台上的结果表明, 所提方法能产生遵循所有约束条件的策略, 利用专家知识显著提高了智能体的安全性.

强化学习在处理复杂多接触机器人操作任务方面显示出巨大的潜力. 然而, 现实世界中需要考虑机器人在训练期间的安全性问题, 以免发生意外碰撞. 为此, Zhu 等[31] 为多接触机器人操作提供了一个接触安全的强化学习框架, 使得机器人在任务空间和关节空间都保持安全. Zhu 等[31] 认为, 关节空间安全不应该仅通过避免碰撞来实现, 而是应该限制接触力的大小. 因此, 当学到的强化学习策略导致机器人手臂与环境之间发生意外碰撞时, 所提框架能够立即检测到碰撞并减小接触力. 同时, 通过强制末端执行器执行多接触任务, 来保持对外部干扰的鲁棒性. 在执行擦拭任务上的真实实验结果表明, 所提框架即使是在策略处于意外碰撞的未知情况下, 仍能在任务空间和关节空间中保持较小的接触力, 同时防止由意外的关节碰撞引起的干扰.

### 4.3 工业过程控制

强化学习在围棋、游戏等领域已取得一定的成功, 这些成功的背后依赖于场景中存在的完美模拟环境. 强化学习算法以不断试错的方式与环境进行交互, 从而获得最优策略. 但是在一些更复杂严苛的现实场景中, 例如复杂工业系统控制, 以不断试错的方式收集数据不仅对成本要求高, 而且可能引发极大的安全隐患. 因此确保系统运行的安全性是进行控制优化的前提. 为此, 针对工业过程控制安全优化问题, 学者们提出了一系列方法.

无模型安全强化学习方法已被应用于化学过程控制领域. 通过迭代建立近似的过程模型, 能够探索控制动作并产生最佳策略. 在此背景下, Savage 等[24] 提出了一种基于高斯过程的 Q 学习算法. 其中, 高斯过程作为一种函数近似方法来描述非等温半间歇反应器的 Q 函数. 通过使用高斯过程对分析不确定性进行编码, 可以在探索与利用之间进行权衡, 从而高效地获得最优策略. 更重要的是, 高斯过程还可以对概率约束违反进行建模, 确保在整个学习过程中进行安全探索. 仿真结果表明, 相比于传统的基于神经网络的强化学习算法, 所提算法仅使用少量的过程轨迹数据就能得到安全有效的控制策略.

强化学习是一种能够处理非线性随机最优控制问题的控制方法. 然而传统的强化学习方法只关注期望回报最大化, 而未考虑满足过程安全约束条件, 因此可能为生化工程带来隐患. 考虑到化学过程控制中的状态约束问题, Pan 等[32] 提出一种 oracle 辅助的约束 Q 学习方法, 用于在随机复杂过程系统中寻找能够满足高概率约束的控制器策略. 具体而言, 在实际奖励函数中添加关于过程约束的惩罚项, 引入约束收紧来限制控制器感知的可行空间, 并使用 Broyden 方法迭代更新回退值, 以惩罚违反约束的行为, 从而保证满足机会约束. 仿真结果表明, 相比于工业上常用的基准控制技术 —— 非线性模型预测控制方法, 所提方法能以更高的概率 (99% 的概率) 保障系统的安全性.

为解决化学过程控制中的模型失配问题, Mowbray 等[25] 采用纯数据驱动的离线安全强化学习方法, 将强化学习策略从离线训练环境 (过程模型) 安全地部署到真实的未知过程模型中. 具体而言, 首先, 使用高斯过程来构建离线的状态空间模型; 其次, 利用相应的后验不确定性预测分布方差来限制期望过程远离约束边界, 从而使模型在安全区域探索; 最后, 采取贝叶斯优化策略来调整约束程度, 从而在性能与风险之间进行更好的平衡. 仿真结果表明该方法能够解释过程的不确定性, 即使在模型失配的情况下也能满足联合机会约束.

优化火力发电机组的燃烧效率是能源行业中一项极具挑战性的任务. 通过充分利用火电行业的历史数据, Zhan 等[22] 将离线强化学习应用于真实的

火力发电燃烧系统中, 并提出一种具有限制性探索的基于模型的离线强化学习 (Model-based offline RL with restrictive exploration, MORE) 算法. 其主要思路为: 首先, 引入额外的成本 critic, 强制模型满足燃烧优化问题的安全约束; 其次, 使用限制性探索方法在不完美的仿真环境中探索样本; 最后, 将探索得到的样本划分为正样本和负样本, 并以混合训练的方式学习安全策略. 在中国 4 个大型燃煤火力发电厂的真实实验表明, 所提算法能够提高火力发电机组的燃烧效率, 降低污染物排放量.

针对信息物理系统控制领域的工业油泵安全控制问题, 赵恒军等[15] 提出了基于增广拉格朗日乘子法的安全强化学习算法. 传统的拉格朗日求解方法要求目标函数具有强凸性和有限性, 严格的收敛条件使得该方法在实际应用中受限. 而增广拉格朗日目标函数放宽了该收敛条件, 进一步提升了算法的鲁棒性. 具体而言, 在原先的拉格朗日目标函数中引入期望损耗的二次惩罚项, 进一步加大损耗惩罚力度. 在增广拉格朗日目标中, 乘子项和二次惩罚项相互协同, 实现惩罚参数的动态调节. 相比于传统基于拉格朗日的安全强化学习方法, 增广拉格朗日方法的收敛条件更具一般性.

### 4.4 电力系统优化

电力系统是安全关键型基础设施的典型例子, 如果违反运行限制则会导致大规模停电, 造成高昂的经济和人力成本. 随着可变的可再生能源资源被整合到电网中, 确保系统状态在运营商定义的 “安全” 区域内已变得越来越重要. 因此, 许多学者将安全强化学习算法应用于解决电力系统优化问题, 以提高训练期间电网的安全性.

针对安全应急电压控制问题, Vu 等[26] 在安全减载的奖励函数中引入障碍函数来指导最优控制策略的搜索. 当系统状态达到安全边界时, 该障碍函数变为负无穷大. 因此, 最大化奖励函数的最优控制策略可以使电力系统避开安全边界, 提高电网在减载期间的安全性.

为辅助孤岛微电网的服务恢复, Du 等[28] 提出了一个两阶段学习框架来确定一个最优的恢复策略, 该框架是建立在基于示范的深度确定性策略梯度算法基础之上. 预训练阶段: 利用模仿学习使智能体具备专家经验, 从而获得较好的初始性能并快速适应环境; 在线训练阶段: 利用动作剪裁、奖励重塑和专家示范来确保安全探索, 同时加速训练过程. 仿真结果表明, 所提策略在处理服务恢复问题上表现出更优的性能, 且比基于模型的方法速度更快, 能应用于实时系统.

针对微电网在线优化调度问题, 季颖和王建辉[16] 提出基于拉格朗日的 SAC 强化学习算法. 将调度问题建模为约束马尔科夫决策过程, 并通过训练卷积神经网络来获得近似最优的在线调度策略. 在真实的电力系统中的仿真结果表明, 在考虑微电网潮流约束情况下, 所提算法仍能学习到最优调度策略.

最近, Tabas 等[33] 将安全强化学习方法用于解决电力系统的频率控制问题, 提出一种基于神经网络的高效安全控制策略. 该策略结合了集合论控制技术与凸优化分析, 并使用简单的线性控制器来寻找最大鲁棒控制不变集, 再利用深度确定性策略梯度算法对神经网络进行端到端训练. 在电力系统中的仿真结果表明, 所提策略确保了安全性且性能优于传统的线性控制器.

### 4.5 医疗健康

近年来, 深度强化学习也被广泛应用于医疗健康领域, 旨在为患者提供更精准有效的治疗. 然而, 安全性是最值得关注的一个问题. 如果处理不当, 则会对患者造成致命伤害. 因此, 有必要为患者设计更为安全可靠的强化学习策略.

针对环境中存在的大规模且复杂的状态空间, 传统的强化学习算法需要智能体进行大量探索才能学习到有效策略. 能否对状态空间进行简化以提高学习效率是一个值得关注的话题. Zhang 等[36] 研究了批处理强化学习 (离线强化学习) 中的时间离散化问题, 对原先的批处理强化学习进行改进. 算法并未将所有状态都视为潜在的决策点, 而只考虑那些具有高行为策略可变性的状态, 即临床医生对病人进行不同治疗的特定决策点, 因此使得状态空间明显缩小. 在一组低血压患者数据中进行验证, 可以发现简化的状态空间能够加快规划速度, 从而辅助临床医生快速诊断.

连续体机械臂已被广泛应用于微创外科手术中. 然而, 由于其非线性行为, 使得这些机器人建模不准确、控制性能差, 如果涉及大脑或心脏等关键器官的手术, 则这些控制错误和不稳定性可能带来致命危害. 为此, Ji 等[21] 提出了一种基于无模型的多智能体深度 Q 网络 (Multi-agent deep Q network, MADQN) 来控制一个自由度为 2 的连续体手术机器人. MADQN 与具有动态改变动作集边界的安全盾 (Shielding) 方法相结合, 有效消除了传统 DQN 框架中离散动作集的限制, 并在瞄准精度和运动稳定性方面提高了连续体机器人的控制性能. 对不同轨迹、外部载荷、软硬扰动以及具有高度

结构非线性的微型机器人的测试实验结果表明, 所提方法能实现亚毫米级的轨迹跟踪精度和较高的稳定性.

传统的深度强化学习无法保证安全性, 因此在机器人辅助微创手术应用中受到限制. 为此, Pore 等[29] 引入一个安全深度强化学习框架, 将安全约束纳入手术子任务的自动化中, 并制定了一个形式化验证工具来评估深度强化学习策略所造成的违规行为. 实验结果表明, 所提框架能在模型执行前识别可能导致安全违规的状态, 从而使机器人能够在安全空间进行操作.

新型冠状病毒肺炎疫情 (COVID-19) 的持续大流行使普通民众的出行受到一定程度的影响. 为规避高风险地区, Misra 等[34] 提出了一种基于 Q 学习的 COVID-19 安全导航系统 S-Nav. 首先, 通过在奖励函数上施加基于区域的惩罚来使智能体学会推荐零风险/低风险的路线; 其次, 为确保推荐的实时性, 利用云计算架构最大限度地减少服务的训练时间和响应时间. 在真实路线图数据上的实验结果显示, S-Nav 几乎能够实时规划出安全路径, 避免出行者交叉感染. 然而, 该系统仅适用于室外环境.

## 5 未来研究方向

尽管近年来安全强化学习领域已涌现出许多新思路新方法, 取得了积极显著的成果, 但仍然有很多问题值得进一步探索. 未来可以从以下几个方面对安全强化学习进行拓展与延伸.

1) 利用基于注意力机制的网络对安全约束进行编码. 更多的强化学习相关研究开始使用基于注意力机制的网络模型, 如 DT[122]、TT[123] 等. 这些网络模型可以挖掘序列间多种维度上的注意力系数, 更好地抽象化路径信息, 提高深度强化学习的样本效率. 同时, 基于注意力机制的网络输入输出灵活性更强, 可以更方便地在编码中嵌入安全信息, 由此进一步增强算法的安全性能.

2) 充分结合已有的强化学习研究成果来提高算法的安全性能. 例如逆强化学习可以扩充经验数据中包含的环境安全信息, 提高安全性评估的精度和网络全局更新的效率; 分层强化学习可以将学习任务分解为多个子任务, 根据子任务设计动态的安全约束和探索方式, 以减小策略陷入局部最优的概率; 元强化学习可以从多个不同的环境中收集安全信息, 使算法学到一个泛化能力更强的安全性评估标准, 从而在没有足够专家经验的下游任务训练中提高安全性能.

3) 将安全约束机制融入离线强化学习中. 当前的离线强化学习算法仅从离线数据中学习, 无需与环境交互, 智能体完全避免了探索步骤, 因此能确保训练期间的安全性. 但如果不对离线强化学习施加额外的安全约束项, 则当策略部署到真实环境时, 无法实现安全保证. 因此, 离线安全强化学习也是未来的研究方向之一.

## 6 结论

深度强化学习是当前机器学习领域的一个热点研究方向, 它为解决复杂的决策与控制问题提供了有效途径. 然而, 大部分强化学习方法并不能直接迁移到真实物理环境中. 传统的强化学习需要智能体不断地与环境交互并以试错的方式来收集数据, 从而获得最优策略, 但这种方式在实际应用中可能会产生巨大开销. 为此, 研究有安全保障的强化学习算法尤为重要. 本文对近年来的安全强化学习算法进行了全面综述与总结, 为研究安全强化学习方向的学者提供指导与思路. 首先对安全强化学习问题进行形式化定义, 将安全强化学习问题转化为约束马尔科夫决策过程, 并总结了安全强化学习常见的两种约束形式. 其次对安全强化学习算法进行分类与汇总, 从修改学习过程、修改学习目标以及离线强化学习三大方面进行综述. 修改学习过程是在智能体的探索过程中施加约束, 因而能同时保证训练时安全和部署时安全. 修改学习目标中拉格朗日法将原先的约束优化问题通过拉格朗日乘子的权衡转化为无约束优化问题, 但这种约束方法属于隐式约束, 在训练期间无法提供安全性保证, 而信赖域法显式地实施安全约束, 利用约束策略优化方法, 确保训练过程中智能体的安全性. 然后介绍了 5 种安全强化学习基准测试平台, 便于研究者进行基准测试和公平比较. 针对修改学习过程和修改学习目标类的安全强化学习方法可以使用 Safety Gym、safe-control-gym 和 SafeRL-Kit 基准测试平台, 而针对离线强化学习, 则使用 D4RL 和 NeoRL 进行测试. 最后总结了安全强化学习在自动驾驶、机器人控制、工业过程控制、电力系统优化和医疗健康 5 大应用领域中的研究进展, 以及展望了未来研究方向. 虽然目前仍有许多问题尚未解决, 但在可预见的未来, 随着学者对安全强化学习问题研究的进一步深入, 相信安全强化学习将成为今后的一个热点话题.

## References

1 Sutton R S, Barto A G. *Reinforcement Learning: An Introduction*. Cambridge: MIT Press, 2018.

2 Dong S, Wang P, Abbas K. A survey on deep learning and its applications. *Computer Science Review*, 2021, **40**: Article No.

100379

3 Wen Zai-Dao, Wang Jia-Rui, Wang Xiao-Xu, Pan Quan. A review of disentangled representation learning. *Acta Automatica Sinica*, 2022, **48**(2): 351−374
(文载道, 王佳蕊, 王小旭, 潘泉. 解耦表征学习综述. 自动化学报, 2022, **48**(2): 351−374)

4 Silver D, Huang A, Maddison C, Guez A, Sifre L, Driessche G, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 2016, **529**(7587): 484−489

5 Shao K, Tang Z T, Zhu Y H, Li N N, Zhao D B. A survey of deep reinforcement learning in video games. arXiv preprint arXiv: 1912.10944, 2019.

6 Kiran B R, Sobh I, Talpaert V, Mannion P, Sallab A A A, Yogamani S, et al. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 2022, **23**(6): 4909−4926

7 Huang Yan-Long, Xu De, Tan Min. On imitation learning of robot movement trajectories: A survey. *Acta Automatica Sinica*, 2022, **48**(2): 315−334
(黄艳龙, 徐德, 谭民. 机器人运动轨迹的模仿学习综述. 自动化学报, 2022, **48**(2): 315−334)

8 Zhang Z D, Zhang D X, Qiu R C. Deep reinforcement learning for power system applications: An overview. *CSEE Journal of Power and Energy Systems*, 2020, **6**(1): 213−225

9 Liu Jian, Gu Yang, Cheng Yu-Hu, Wang Xue-Song. Prediction of breast cancer pathogenic genes based on multi-agent reinforcement learning. *Acta Automatica Sinica*, 2022, **48**(5): 1246−1258
(刘健, 顾扬, 程玉虎, 王雪松. 基于多智能体强化学习的乳腺癌致病基因预测. 自动化学报, 2022, **48**(5): 1246−1258)

10 García J, Fernández F. A comprehensive survey on safe reinforcement learning. *The Journal of Machine Learning Research*, 2015, **16**(1): 1437−1480

11 Altman E. *Constrained Markov Decision Processes: Stochastic Modeling*. New York: Routledge, 1999.

12 Kamran D, Ren Y, Lauer M. High-level decisions from a safe maneuver catalog with reinforcement learning for safe and cooperative automated merging. In: Proceedings of the IEEE International Intelligent Transportation Systems Conference (ITSC). Indiana, USA: IEEE, 2021. 804−811

13 Trumpp R, Bayerlein H, Gesbert D. Modeling interactions of autonomous vehicles and pedestrians with deep multi-agent reinforcement learning for collision avoidance. In: Proceedings of the IEEE Intelligent Vehicles Symposium (IV). Aachen, Germany: IEEE, 2022. 331−336

14 Yang T Y, Zhang T N, Luu L, Ha S, Tan J, Yu W H. Safe reinforcement learning for legged locomotion. arXiv preprint arXiv: 2203.02638, 2022.

15 Zhao Heng-Jun, Li Quan-Zhong, Zeng Xia, Liu Zhi-Ming. Safe reinforcement learning algorithm and its application in intelligent control for CPS. *Journal of Software*, 2022, **33**(7): 2538−2561
(赵恒军, 李权忠, 曾霞, 刘志明. 安全强化学习算法及其在 CPS 智能控制中的应用. 软件学报, 2022, **33**(7): 2538−2561)

16 Ji Ying, Wang Jian-Hui. Online optimal scheduling of a microgrid based on deep reinforcement learning. *Control and Decision*, 2022, **37**(7): 1675−1684
(季颖, 王建辉. 基于深度强化学习的微电网在线优化调度. 控制与决策, 2022, **37**(7): 1675−1684)

17 Zhang L R, Zhang Q, Shen L, Yuan B, Wang X Q. SafeRL-Kit: Evaluating efficient reinforcement learning methods for safe autonomous driving. arXiv preprint arXiv: 2206.08528, 2022.

18 Thananjeyan B, Balakrishna A, Nair S, Luo M, Srinivasan K, Hwang M, et al. Recovery RL: Safe reinforcement learning with learned recovery zones. *IEEE Robotics and Automation Letters*, 2021, **6**(3): 4915−4922

19 Levine S, Kumar A, Tucker G, Fu J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. arXiv preprint arXiv: 2005.01643, 2020.

20 Prudencio R F, Máximo M R O A, Colombini E L. A survey on offline reinforcement learning: Taxonomy, review, and open problems. arXiv preprint arXiv: 2203.01387, 2022.

21 Ji G L, Yan J Y, Du J X, Yan W Q, Chen J B, Lu Y K, et al. Towards safe control of continuum manipulator using shielded multiagent reinforcement learning. *IEEE Robotics and Automation Letters*, 2021, **6**(4): 7461−7468

22 Zhan X Y, Xu H R, Zhang Y, Zhu X Y, Yin H L, Zheng Y. DeepThermal: Combustion optimization for thermal power generating units using offline reinforcement learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. California, USA: AAAI Press, 2022. 4680−4688

23 Zhang Y X, Gao B Z, Guo L L, Guo H Y, Chen H. Adaptive decision-making for automated vehicles under roundabout scenarios using optimization embedded reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2021, **32**(12): 5526−5538

24 Savage T, Zhang D D, Mowbray M, Chanona E A D R. Model-free safe reinforcement learning for chemical processes using Gaussian processes. *IFAC-PapersOnLine*, 2021, **54**(3): 504−509

25 Mowbray M, Petsagkourakis P, Chanona E A D R, Zhang D D. Safe chance constrained reinforcement learning for batch process control. *Computers & Chemical Engineering*, 2022, **157**: Article No. 107630

26 Vu T L, Mukherjee S, Huang R K, Huang Q H. Barrier function-based safe reinforcement learning for emergency control of power systems. In: Proceedings of the 60th IEEE Conference on Decision and Control (CDC). Texas, USA: IEEE, 2021. 3652−3657

27 García J, Shafie D. Teaching a humanoid robot to walk faster through safe reinforcement learning. *Engineering Applications of Artificial Intelligence*, 2020, **88**: Article No. 103360

28 Du Y, Wu D. Deep reinforcement learning from demonstrations to assist service restoration in islanded microgrids. *IEEE Transactions on Sustainable Energy*, 2022, **13**(2): 1062−1072

29 Pore A, Corsi D, Marchesini E, Dall′Alba D, Casals A, Farinelli A, et al. Safe reinforcement learning using formal verification for tissue retraction in autonomous robotic-assisted surgery. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Prague, Czech Republic: IEEE, 2021. 4025−4031

30 Dai Shan-Shan, Liu Quan. Action constrained deep reinforcement learning based safe automatic driving method. *Computer Science*, 2021, **48**(9): 235−243
(代珊珊, 刘全. 基于动作约束深度强化学习的安全自动驾驶方法. 计算机科学, 2021, **48**(9): 235−243)

31 Zhu X, Kang S C, Chen J Y. A contact-safe reinforcement learning framework for contact-rich robot manipulation. In: Proceedings of the International Conference on Intelligent Robots and Systems (IROS). Kyoto, Japan: IEEE, 2022. 2476−2482

32 Pan E, Petsagkourakis P, Mowbray M, Zhang D D, Chanona E A D R. Constrained model-free reinforcement learning for process optimization. *Computers & Chemical Engineering*, 2021, **154**: Article No. 107462

33 Tabas D, Zhang B S. Computationally efficient safe reinforcement learning for power systems. In: Proceedings of the American Control Conference. Georgia, USA: IEEE, 2022. 3303−3310

34 Misra S, Deb P K, Koppala N, Mukherjee A, Mao S W. S-Nav: Safety-aware IoT navigation tool for avoiding COVID-19 hotspots. *IEEE Internet of Things Journal*, 2021, **8**(8): 6975−6982

35 Corsi D, Yerushalmi R, Amir G, Farinelli A, Harel D, Katz G. Constrained reinforcement learning for robotics via scenario-

based programming. arXiv preprint arXiv: 2206.09603, 2022.

36 Zhang K, Wang Y H, Du J Z, Chu B, Celi L A, Kindle R, et al. Identifying decision points for safe and interpretable reinforcement learning in hypotension treatment. arXiv preprint arXiv: 2101.03309, 2021.

37 Zhao T Z, Luo J L, Sushkov O, Pevceviciute R, Heess N, Scholz J, et al. Offline meta-reinforcement learning for industrial insertion. In: Proceedings of the International Conference on Robotics and Automation (ICRA). Philadelphia, PA, USA: IEEE, 2022. 6386−6393

38 Sui Y N, Gotovos A, Burdick J W, Krause A. Safe exploration for optimization with Gaussian processes. In: Proceedings of the International Conference on Machine Learning. Lille, France: PMLR, 2015. 997−1005

39 Turchetta M, Berkenkamp F, Krause A. Safe exploration in finite Markov decision processes with Gaussian processes. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: Curran Associates Inc., 2016. 4312−4320

40 Wachi A, Kajino H, Munawar A. Safe exploration in Markov decision processes with time-variant safety using spatio-temporal Gaussian process. arXiv preprint arXiv: 1809.04232, 2018.

41 Alshiekh M, Bloem R, Ehlers R, Könighofer B, Niekum S, Topcu U. Safe reinforcement learning via shielding. In: Proceedings of the AAAI Conference on Artificial Intelligence. Lousiana, USA: AAAI Press, 2018. 2669−2678

42 Zhang W B, Bastani O, Kumar V. MAMPS: Safe multi-agent reinforcement learning via model predictive shielding. arXiv preprint arXiv: 1910.12639, 2019.

43 Jansen N, Könighofer B, Junges S, Serban A C, Bloem R. Safe reinforcement learning via probabilistic shields. arXiv preprint arXiv: 1807.06096, 2018.

44 Li S, Bastani O. Robust model predictive shielding for safe reinforcement learning with stochastic dynamics. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Paris, France: IEEE, 2020. 7166−7172

45 Bastani O. Safe reinforcement learning with nonlinear dynamics via model predictive shielding. In: Proceedings of the American Control Conference. Los Angeles, USA: IEEE, 2021. 3488−3494

46 Perkins T J, Barto A G. Lyapunov design for safe reinforcement learning. *Journal of Machine Learning Research*, 2003, **3**: 803−832

47 Berkenkamp F, Turchetta M, Schoellig A, Krause A. Safe model-based reinforcement learning with stability guarantees. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. California, USA: Curran Associates Inc., 2017. 908−919

48 Chow Y, Nachum O, Faust A, Ghavamzadeh M, Duéñez-Guzmán E. Lyapunov-based safe policy optimization for continuous control. arXiv preprint arXiv: 1901.10031, 2019.

49 Jeddi A B, Dehghani N L, Shafieezadeh A. Lyapunov-based uncertainty-aware safe reinforcement learning. arXiv preprint arXiv: 2107.13944, 2021.

50 Cheng R, Orosz G, Murray R M, Burdick J W. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. In: Proceedings of the AAAI Conference on Artificial Intelligence. Hawaii, USA: AAAI Press, 2019. 3387−3395

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

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

53 Emam Y, Notomista G, Glotfelter P, Kira Z, Egerstedt M. Safe model-based reinforcement learning using robust control barrier functions. arXiv preprint arXiv: 2110.05415, 2021.

54 Bura A, HasanzadeZonuzy A, Kalathil D, Shakkottai S, Chamberland J F. Safe exploration for constrained reinforcement learning with provable guarantees. arXiv preprint arXiv: 2112.00885, 2021.

55 Thomas G, Luo Y P, Ma T Y. Safe reinforcement learning by imagining the near future. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. Montreal, Canada: Curran Associates Inc., 2021. 13859−13869

56 Ma Y J, Shen A, Bastani O, Jayaraman D. Conservative and adaptive penalty for model-based safe reinforcement learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press, 2022. 5404−5412

57 Saunders W, Sastry G, Stuhlmüller A, Evans O. Trial without error: Towards safe reinforcement learning via human intervention. In: Proceedings of the International Conference on Autonomous Agents and MultiAgent Systems. Stockholm, Sweden: IFAAMAS, 2018. 2067−2069

58 Prakash B, Khatwani M, Waytowich N, Mohsenin T. Improving safety in reinforcement learning using model-based architectures and human intervention. In: Proceedings of the International Flairs Conference. Florida, USA: AAAI Press, 2019. 50−55

59 Sun H, Xu Z P, Fang M, Peng Z H, Guo J D, Dai B, et al. Safe exploration by solving early terminated MDP. arXiv preprint arXiv: 2107.04200, 2021.

60 Prakash B, Waytowich N R, Ganesan A, Oates T, Mohsenin T. Guiding safe reinforcement learning policies using structured language constraints. In: Proceedings of the SafeAI Workshop of AAAI Conference on Artificial Intelligence. New York, USA: AAAI Press, 2020. 153−161

61 Yang T Y, Hu M, Chow Y, Ramadge P J, Narasimhan K. Safe reinforcement learning with natural language constraints. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. Montreal, Canada: Curran Associates Inc., 2021. 13794−13808

62 Turchetta M, Kolobov A, Shah S, Krause A, Agarwal A. Safe reinforcement learning via curriculum induction. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2020. 12151−12162

63 Peng Z H, Li Q Y, Liu C X, Zhou B L. Safe driving via expert guided policy optimization. In: Proceedings of the 5th Conference on Robot Learning. London, UK: PMLR, 2022. 1554−1563

64 Li Q Y, Peng Z H, Zhou B L. Efficient learning of safe driving policy via human-AI copilot optimization. arXiv preprint arXiv: 2202.10341, 2022.

65 Dalal G, Dvijotham K, Vecerik M, Hester T, Paduraru C, Tassa Y. Safe exploration in continuous action spaces. arXiv preprint arXiv: 1801.08757, 2018.

66 Zhu Fei, Wu Wen, Fu Yu-Chen, Liu Quan. A dual deep network based secure deep reinforcement learning method. *Chinese Journal of Computers*, 2019, **42**(8): 1812−1826
(朱斐, 吴文, 伏玉琛, 刘全. 基于双深度网络的安全深度强化学习方法. 计算机学报, 2019, **42**(8): 1812−1826)

67 Zheng L Y, Shi Y Y, Ratliff L J, Zhang B. Safe reinforcement learning of control-affine systems with vertex networks. In: Proceedings of the 3rd Conference on Learning for Dynamics and Control. Zurich, Switzerland: PMLR, 2021. 336−347

68 Marchesini E, Corsi D, Farinelli A. Exploring safer behaviors for deep reinforcement learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press, 2022. 7701−7709

69 Mannucci T, van Kampen E J, de Visser C, Chu Q P. Safe exploration algorithms for reinforcement learning controllers. *IEEE Transactions on Neural Networks and Learning Systems*,

2018, **29**(4): 1069−1081

70 Memarzadeh M, Pozzi M. Model-free reinforcement learning with model-based safe exploration: Optimizing adaptive recovery process of infrastructure systems. *Structural Safety*, 2019, **80**: 46−55

71 Wachi A, Wei Y Y, Sui Y N. Safe policy optimization with local generalized linear function approximations. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. Montreal, Canada: Curran Associates Inc., 2021. 20759−20771

72 Chow Y, Ghavamzadeh M, Janson L, Pavone M. Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research*, 2017, **18**(1): 6070−6120

73 Ma H T, Guan Y, Li S E, Zhang X T, Zheng S F, Chen J Y. Feasible actor-critic: Constrained reinforcement learning for ensuring statewise safety. arXiv preprint arXiv: 2105.10682, 2021.

74 Roy J, Girgis R, Romoff J, Bacon P L, Pal C. Direct behavior specification via constrained reinforcement learning. In: Proceedings of the International Conference on Machine Learning. Maryland, USA: PMLR, 2022. 18828−18843

75 Sootla A, Cowen-Rivers A I, Jafferjee T, Wang Z Y, Mguni D H, Wang J, et al. Sauté RL: Almost surely safe reinforcement learning using state augmentation. In: Proceedings of the International Conference on Machine Learning. Maryland, USA: PMLR, 2022. 20423−20443

76 Tessler C, Mankowitz D J, Mannor S. Reward constrained policy optimization. arXiv preprint arXiv: 1805.11074, 2018.

77 Yu M, Yang Z R, Kolar M, Wang Z R. Convergent policy optimization for safe reinforcement learning. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2019. 3127−3139

78 Bai Q B, Bedi A S, Agarwal M, Koppel A, Aggarwal V. Achieving zero constraint violation for constrained reinforcement learning via primal-dual approach. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press, 2022. 3682−3689

79 Achiam J, Held D, Tamar A, Abbeel P. Constrained policy optimization. In: Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: PMLR, 2017. 22−31

80 Schulman J, Levine S, Moritz P, Jordan M, Abbeel P. Trust region policy optimization. In: Proceedings of the International Conference on Machine Learning. Lille, France: PMLR, 2015. 1889−1897

81 Yang T Y, Rosca J, Narasimhan K, Ramadge P J. Projection-based constrained policy optimization. arXiv preprint arXiv: 2010.03152, 2020.

82 Zhang Y M, Vuong Q, Ross K W. First order constrained optimization in policy space. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2020. 15338−15349

83 Zhang L R, Shen L, Yang L, Chen S X, Yuan B, Wang X Q, et al. Penalized proximal policy optimization for safe reinforcement learning. arXiv preprint arXiv: 2205.11814, 2022.

84 Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. arXiv preprint arXiv: 1707.06347, 2017.

85 Xu T Y, Liang Y B, Lan G H. CRPO: A new approach for safe reinforcement learning with convergence guarantee. In: Proceedings of the International Conference on Machine Learning. Vienna, Austria: PMLR, 2021. 11480−11491

86 Liu Z X, Cen Z P, Isenbaev V, Liu W, Wu Z S, Li B, et al. Constrained variational policy optimization for safe reinforcement learning. In: Proceedings of the International Conference on Machine Learning. Maryland, USA: PMLR, 2022. 13644−13668

87 Fujimoto S, Meger D, Precup D. Off-policy deep reinforcement learning without exploration. In: Proceedings of the International Conference on Machine Learning. California, USA: PMLR, 2019. 2052−2062

88 Kumar A, Fu J, Soh M, Tucker G, Levine S. Stabilizing off-policy Q-learning via bootstrapping error reduction. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2019. 11784−11794

89 Zhou W X, Bajracharya S, Held D. PLAS: Latent action space for offline reinforcement learning. In: Proceedings of the Conference on Robot Learning. Cambridge, USA: PMLR, 2020. 1719−1735

90 Chen X, Ghadirzadeh A, Yu T H, Gao Y, Wang J H, Li W Z, et al. Latent-variable advantage-weighted policy optimization for offline RL. arXiv preprint arXiv: 2203.08949, 2022.

91 Kumar A, Zhou A, Tucker G, Levine S. Conservative Q-learning for offline reinforcement learning. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2020. 1179−1191

92 Xu H R, Zhan X Y, Zhu X Y. Constraints penalized Q-learning for safe offline reinforcement learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. California, USA: AAAI Press, 2022. 8753−8760

93 Kostrikov I, Nair A, Levine S. Offline reinforcement learning with implicit Q-learning. arXiv preprint arXiv: 2110.06169, 2021.

94 Zhang R Y, Dai B, Li L H, Schuurmans D. GenDICE: Generalized offline estimation of stationary values. arXiv preprint arXiv: 2002.09072, 2020.

95 Zhan W H, Huang B H, Huang A, Jiang N, Lee J D. Offline reinforcement learning with realizability and single-policy concentrability. In: Proceedings of the Conference on Learning Theory. London, UK: PMLR, 2022. 2730−2775

96 Siegel N Y, Springenberg J T, Berkenkamp F, Abdolmaleki A, Neunert M, Lampe T, et al. Keep doing what worked: Behavioral modelling priors for offline reinforcement learning. arXiv preprint arXiv: 2002.08396, 2020.

97 Wang Z Y, Novikov A, Zolna K, Springenberg J T, Reed S, Shahriari B, et al. Critic regularized regression. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2020. 7768−7778

98 Emmons S, Eysenbach B, Kostrikov I, Levine S. RvS: What is essential for offline RL via supervised learning. arXiv preprint arXiv: 2112.10751, 2021.

99 Uchendu I, Xiao T, Lu Y, Zhu B H, Yan M Y, Simon J, et al. Jump-start reinforcement learning. arXiv preprint arXiv: 2204.02372, 2022.

100 Ray A, Achiam J, Amodei D. Benchmarking safe exploration in deep reinforcement learning. arXiv preprint arXiv: 1910.01708, 2019.

101 Hawkins D. *Constrained Optimization and Lagrange Multiplier Methods*. Boston: Academic Press, 1982.

102 Yuan Z C, Hall A W, Zhou S Q, Brunke L, Greeff M, Panerati J, et al. Safe-control-gym: A unified benchmark suite for safe learning-based control and reinforcement learning. arXiv preprint arXiv: 2109.06325, 2021.

103 Buchli J, Farshidian F, Winkler A, Sandy T, Giftthaler M. Optimal and learning control for autonomous robots. arXiv preprint arXiv: 1708.09342, 2017.

104 Rawlings J B, Mayne D Q, Diehl M M. *Model Predictive Control: Theory, Computation, and Design*. Madison, Wisconsi: Nob Hill Publishing, 2017.

105 Haarnoja T, Zhou A, Abbeel P, Levine S. Soft actor-critic: Off-

policy maximum entropy deep reinforcement learning with a stochastic actor. In: Proceedings of the International Conference on Machine Learning. Stockholm, Sweden: PMLR, 2018. 1861−1870

106 Hewing L, Kabzan J, Zeilinger M N. Cautious model predictive control using Gaussian process regression. *IEEE Transactions on Control Systems Technology*, 2020, **28**(6): 2736−2743

107 Pinto L, Davidson J, Sukthankar R, Gupta A. Robust adversarial reinforcement learning. In: Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: PMLR, 2017. 2817−2826

108 Vinitsky E, Du Y Q, Parvate K, Jang K, Abbeel P, Bayen A. Robust reinforcement learning using adversarial populations. arXiv preprint arXiv: 2008.01825, 2020.

109 Wabersich K P, Zeilinger M N. Linear model predictive safety certification for learning-based control. In: Proceedings of the IEEE Conference on Decision and Control (CDC). Florida, USA: IEEE, 2018. 7130−7135

110 Ames A D, Coogan S, Egerstedt M, Notomista G, Sreenath K, Tabuada P. Control barrier functions: Theory and applications. In: Proceedings of the 18th European Control Conference (ECC). Naples, Italy: IEEE, 2019. 3420−3431

111 Yang L, Ji L M, Dai J T, Zhang L R, Zhou B B, Li P F, et al. Constrained update projection approach to safe policy optimization. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. 9111−9124

112 Li Q Y, Peng Z H, Feng L, Zhang Q H, Xue Z H, Zhou B L. MetaDrive: Composing diverse driving scenarios for generalizable reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, **45**(3): 3461−3475

113 Ha S, Xu P, Tan Z Y, Levine S, Tan J. Learning to walk in the real world with minimal human effort. arXiv preprint arXiv: 2002.08550, 2020.

114 Fu J, Kumar A, Nachum O, Tucker G, Levine S. D4RL: Datasets for deep data-driven reinforcement learning. arXiv preprint arXiv: 2004.07219, 2020.

115 Wu Y F, Tucker G, Nachum O. Behavior regularized offline reinforcement learning. arXiv preprint arXiv: 1911.11361, 2019.

116 Peng X B, Kumar A, Zhang G, Levine S. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. arXiv preprint arXiv: 1910.00177, 2019.

117 Nachum O, Dai B, Kostrikov I, Chow Y, Li L H, Schuurmans D. AlgaeDICE: Policy gradient from arbitrary experience. arXiv preprint arXiv: 1912.02074, 2019.

118 Qin R J, Gao S Y, Zhang X Y, Xu Z, Huang S K, Li Z W, et al. NeoRL: A near real-world benchmark for offline reinforcement learning. arXiv preprint arXiv: 2102.00714, 2021.

119 Matsushima T, Furuta H, Matsuo Y, Nachum O, Gu S X. Deployment-efficient reinforcement learning via model-based offline optimization. arXiv preprint arXiv: 2006.03647, 2020.

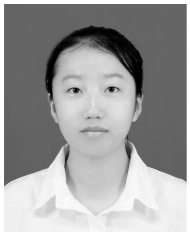
120 Yu T H, Thomas G, Yu L T, Ermon S, Zou J, Levine S, et al. MOPO: Model-based offline policy optimization. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2020. 14129−14142

121 Brunke L, Greeff M, Hall A W, Yuan Z C, Zhou S Q, Panerati J, et al. Safe learning in robotics: From learning-based control to safe reinforcement learning. arXiv preprint arXiv: 2108.06266, 2021.

122 Chen L L, Lu K, Rajeswaran A, Lee K, Grover A, Laskin M, et al. Decision transformer: Reinforcement learning via sequence modeling. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. Sydney, Australia: Curran Associates Inc., 2021. 15084−15097

123 Janner M, Li Q Y, Levine S. Offline reinforcement learning as one big sequence modeling problem. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. Sydney, Australia: Curran Associates Inc., 2021. 1273−1286

**王雪松** 中国矿业大学教授. 2002 年获得中国矿业大学博士学位. 主要研究方向为机器学习, 模式识别.
E-mail: wangxuesongcumt@163.com
(**WANG Xue-Song** Professor at China University of Mining and Technology. She received her Ph.D. degree from China University of Mining and Technology in 2002. Her research interest covers machine learning and pattern recognition.)



**王荣荣** 中国矿业大学博士研究生. 2021 年获得济南大学硕士学位. 主要研究方向为深度强化学习.
E-mail: wangrongrong1996@126.com
(**WANG Rong-Rong** Ph.D. candidate at China University of Mining and Technology. She received her master degree from University of Jinan in 2021. Her main research interest is deep reinforcement learning.)

**程玉虎** 中国矿业大学教授. 2005 年获得中国科学院自动化研究所博士学位. 主要研究方向为机器学习, 智能系统. 本文通信作者.
E-mail: chengyuhu@163.com
(**CHENG Yu-Hu** Professor at China University of Mining and Technology. He received his Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences in 2005. His research interest covers machine learning and intelligent system. Corresponding author of this paper.)