安全强化学习及其在机器人系统中的应用综述

张昌昕, 张兴龙, 徐　昕†, 陆　阳

(国防科技大学 智能科学学院, 湖南 长沙 410000)

摘要: 强化学习是一类通过与环境交互实现序贯优化决策的机器学习方法, 已经在游戏、推荐系统及自然语言处理等任务中得到了应用. 然而, 强化学习算法应用于真实世界中的机器人系统时, 如何保证安全性仍然面临挑战. 近年来, 针对机器人系统的安全强化学习方法研究已经成为热点方向, 获得了机器人和强化学习领域的广泛关注. 本文结合现有的工作, 综述了安全强化学习理论和方法的重要成果和发展趋势, 并重点关注了现有方法在机器人领域的适用性. 本文首先给出了安全强化学习的一般问题描述. 其次, 从方法和性能的角度重点介绍了该领域的最新重要进展, 包括约束策略优化、控制障碍函数、安全过滤器和对抗性博弈训练等方法, 以及安全强化学习方法在地面移动机器人系统、无人飞行器和其他机器人系统中的应用情况. 最后, 对该领域的未来研究方向进行了展望和探讨.

关键词: 机器人; 安全强化学习; 约束马尔可夫决策过程; 鲁棒性

引用格式: 张昌昕, 张兴龙, 徐昕, 等. 安全强化学习及其在机器人系统中的应用综述. 控制理论与应用, 2023, 40(12): 2090 – 2103

DOI: 10.7641/CTA.2023.30247

Safe reinforcement learning and its applications in robotics: A survey

ZHANG Chang-xin, ZHANG Xing-long, XU Xin†, LU Yang

(College of Intelligence Science and Technology, National University of Defense Technology, Changsha Hunan 410000, China)

Abstract: Reinforcement learning is a kind of machine learning method that realizes sequential optimization decisions by interacting with the environment. It has been applied in games, recommendation systems and natural language processing. However, it is still a challenge to ensure the safety of reinforcement learning algorithms when applied to robotics in the real world. In recent years, the safe reinforcement learning methods for robotics systems have become a hot research direction, gaining extensive attention in robotics and reinforcement learning communities. This paper surveys important achievements and development tendency of safe reinforcement learning based on the existing work and focuses on their applicability in robotics. This paper first introduces the general problem description of safe reinforcement learning. Then we focus on the latest significant progress in this field from the perspective of method and performance, including constraint policy optimization, control barrier function, safety filter and adversarial training methods, and their applications in autonomous driving vehicles, unmanned aerial vehicles and other robotic systems. Finally, the future research direction of this field is prospected and discussed.

Key words: robotics; safe reinforcement learning; constrained Markov decision process; robustness

Citation: ZHANG Changxin, ZHANG Xinglong, XU Xin, et al. Safe reinforcement learning and its applications in robotics: A survey. *Control Theory & Applications*, 2023, 40(12): 2090 – 2103

1　引言

作为机械化、信息化和智能化融合发展的重要体现, 机器人系统在人类社会的生产生活中扮演着越来越重要的角色[1–3]. 地面移动机器人系统[4]、无人飞行器[5]及仿生机器人[6]等机器人系统将很大程度上方便人类的生活并提高生产效率. 然而, 这些机器人系统往往需要在复杂不确定场景中实现智能感知与优化决策[7]. 为了解决这一问题, 研究人员通常采用机器学习方法, 使得机器人系统利用经验数据来学习知识并优化性能[8–9]. 在此背景下, 由于机器人系统本身及其面临的环境日益复杂, 如何保证其安全有效地学习成为一个重要的需求和挑战.

强化学习(reinforcement learning, RL)是一类通过与环境交互实现序贯优化决策的机器学习方法, 以探

收稿日期: 2023−04−23; 录用日期: 2023−11−28.

†通信作者. E-mail: xuxin_mail@263.net.

本文责任编委: 辛景民.

国家自然科学基金项目(62003361, U21A20518)资助.

Supported by the National Natural Science Foundation of China (62003361, U21A20518).

索环境并最大化累积回报的方式使智能体能够在未知环境中做出自主决策[10]. 近年来, 针对强化学习中的高维连续特征表示问题, 深度强化学习方法[11] (deep RL, DRL)已经成为该领域的一个重要方向和趋势. DRL方法集成了深度学习 (deep learning, DL)[12]在特征表达上的优势和强化学习在优化决策方面的能力, 目前已经在许多领域开展广泛的应用研究, 如游戏[13]、推荐系统[14]、自然语言处理[15]、医疗健康[16]和电网优化调度[17]等. 由于机器人系统通常需要与环境交互并学习行为策略, 使得强化学习在智能机器人系统中的应用得到广泛关注与研究[18–21]. 但由于环境的不确定性和复杂性, 以及在策略学习过程中需要对未知环境进行探索, 机器人系统可能会采取不安全的动作, 导致意外事故的发生. 在某些任务中, 机器人系统的安全性至关重要[22], 例如自动驾驶车辆的危险行为可能造成的损失不可忽视, 甚至可能威胁到乘客的生命安全, 因此, 如何确保强化学习在实际应用中安全可靠是一个关键问题. 强化学习方法应用于机器人系统时, 机器人系统本身具有较强的不确定性和高动态性, 并且在现实世界中面临的环境更加复杂. 这些特点给机器人系统的安全性带来了诸多挑战. 因此, 针对机器人系统, 设计具有安全性保证的强化学习方法受到越来越广泛的关注.

安全强化学习 (safe RL, SRL) 可以定义为确保系统表现合理或遵守安全约束的强化学习, 旨在提高传统强化学习算法的安全性. Garcia和Fern[23]将安全强化学习分为两类方法: 一类为基于风险考量修改优化准则的方法; 另一类为基于外部知识修改探索过程的方法. 在 Gu 等[24]的论文中, 根据是否利用模型信息将安全强化学习方法分为基于模型和无模型两类. Brunke等[7]则从强化学习和控制方法两个角度阐述了机器人系统的安全学习控制. 此外, 王雪松等[25]将离线强化学习方法归类为安全强化学习方法, 这种方法仅通过离线经验数据来学习最优策略, 避免了在环境中进行探索的过程. 近年来, 安全强化学习领域涌现出多种多样的方法, 几年时间内相关文献数量也从几十篇增长到数百篇[7].

本文结合现有的工作回顾了安全强化学习的一些重要进展和成果, 并重点关注了这些方法在机器人领域的应用研究. 在本文中, 将安全强化学习方法分为满足安全约束与增强策略鲁棒性两类方法, 这两类方法分别解决两种不同类型的安全问题, 这种分类方式提供了与以往综述文献不同的视角来分析安全强化学习的方法脉络. 在本文中, 首先介绍了安全强化学习的一般问题描述: 强化学习问题可以建模为马尔可夫决策过程(Markov decision process, MDP)[26]. 为了对智能体系统进行安全约束, 可以将传统MDP扩展为约束MDP(constrained MDP, CMDP)[22], 在优化问题中引入约束; 另外, 受鲁棒控制的启发, 可以将MDP扩展为鲁棒MDP(robust MDP, RMDP)[27], 旨在提高机器人系统对不确定因素或风险的鲁棒性. 然后, 本文对安全强化学习的重要算法与理论研究进展进行了分类分析和讨论, 包括约束策略优化、控制障碍函数、安全过滤器和对抗性博弈训练等方法. 继而介绍了这类方法在地面移动机器人系统和无人飞行器等机器人系统中的应用研究情况. 最后, 对安全强化学习方法及其在机器人系统中应用的未来研究方向进行了展望和探讨.

本文组织结构如下: 第2节对安全强化学习问题进行了分类和形式化描述; 第3节对安全强化学习方法的重要算法与理论研究进展进行了分类讨论与综述; 第4 节介绍了安全强化学习在机器人领域的应用研究情况; 第5节探讨了安全强化学习及在机器人系统中应用的未来研究方向; 第6节对全文进行总结.

2　问题描述

安全强化学习问题可以视作对强化学习问题的扩展, 旨在在学习或者部署过程中保证智能体以及周围环境的安全性. 本节介绍了强化学习问题在安全性方面的两类扩展, 分别为考虑安全约束的强化学习问题和考虑鲁棒性的强化学习问题.

强化学习问题通常由数学理想化形式的马尔可夫决策过程(MDP)进行建模, 用于描述序列决策问题. MDP中包含了环境状态(state), 动作(action)以及奖励(reward)等元素, 可以表示为元组$\langle S, A, P, R, \gamma\rangle$, 其中: S为环境状态空间; A为智能体可以执行的动作空间; $P: S\times A\times S\to[0,1]$为环境状态转移函数; $R: S\times A\times S\to\mathbb{R}$为奖励函数; $\gamma\in[0,1]$ 为折扣因子, 在计算长期回报过程中决定着智能体对当前和未来奖励的重视程度. 在MDP中, 智能体在t时刻使用策略(policy)π, 根据当前状态$s_t\in S$选择动作$a_t\in A$并执行, 然后通过环境状态转移函数P和奖励函数R得到下一时刻环境状态s_{t+1}并获得奖励r_t. 在这些经验数据的基础上, 智能体根据得到的奖励反馈对策略π进行调整, 通过最大化累积回报来提升行为策略的性能, 构成了学习过程. 在计算累计回报时, 通常使用折扣因子γ对离当前状态更远的未来回报进行折衰减. 该期望累积回报可以表示为策略π下的状态值函数, 即

$$V^{\pi}(s)=\mathrm{E}_{\pi}[\sum_{t=0}^{\infty}\gamma^{t}R(s_t,a_t,s_{t+1})\mid s_0=s],\quad(1)$$

其中$\mathrm{E}_{\pi}[\cdot]$表示数学期望. 在强化学习中, 智能体可以通过经验数据学习值函数并对策略进行评估与改进, 最终得到最优的值函数和行为策略.

2.1　考虑安全约束的RL问题

在强化学习中, 可以通过设置一些安全约束来定

义和量化智能体的安全性. 这些安全约束通常基于物理限制、法律法规或道德准则等来制定, 然后将其纳入学习过程以监控和调整智能体的行为. 为了对安全约束进行描述, 一种常用的建模方式为约束马尔可夫决策过程(CMDP)[28]. CMDP是标准马尔可夫决策过程的扩展形式, 可以通过元组$\langle S, A, P, R, \gamma, C\rangle$来表示, 其中$C$表示约束集(constraint set). 定义$n_{\rm c}$个状态约束函数: $c_t(s_t, a_t, s_{t+1}) \in \mathbb{R}^{n_{\rm c}}$, 每个约束$c_t^j$为实值的时变函数. 轨迹约束可以定义为从初始时刻状态s_0开始, 经过一系列动作和状态, 对轨迹的累积成本进行约束[24], 即

$$V_{c_t^j}^{\pi} = \mathrm{E}_{\pi}[\sum_{t=0}^{\infty} \gamma^t c_t^j(s_t, a_t, s_{t+1})] \leqslant d_j, \tag{2}$$

其中: $V_{c_t^j}^{\pi}$为策略π下状态动作轨迹中违反约束累积折扣成本的期望, d_j为$V_{c_t^j}^{\pi}$对应的约束阈值. 根据文献[24], 累计成本约束(2)还有两个变式, 即成本均值约束

$$V_{c_t^j}^{\pi} = \mathrm{E}_{\pi}[\frac{1}{T}\sum_{t=0}^{T-1} \gamma^t c_t^j(s_t, a_t, s_{t+1})] \leqslant d_j \tag{3}$$

和成本概率约束

$$V_{c_t^j}^{\pi} = \Pr(\sum_{t} c_t^j(s_t, a_t, s_{t+1}) \geqslant \eta) \leqslant d_j, \tag{4}$$

其中: $\Pr(\cdot)$表示概率, η为常数.

轨迹约束的优点是可以容易地描述并解决约束优化问题, 缺点是难以在学习过程中对轨迹上每一时刻的状态和动作进行约束, 以保证每一时刻的安全性[24, 29]. 为了对系统的学习或部署过程中的每一时刻的状态和动作施加约束, 可以引入状态约束. 状态约束按照严格程度从强到弱划分为硬约束(hard constraints)、概率约束(probabilistic constraints)和软约束(soft constraints)[7]:

1) 硬约束. 在所有时间里$t \in \{0, \cdots, N\}$, 系统满足硬约束

$$c_t^j(s_t, a_t, s_{t+1}) \leqslant 0; \tag{5}$$

2) 概率约束. 系统满足约束

$$\Pr(c_t^j(s_t, a_t, s_{t+1}) \leqslant 0) \geqslant p^j, \tag{6}$$

其中$p^j \in (0, 1)$定义了第j个约束满足的可能性(likelihood), 当$p^j = 1$时, 概率约束与硬约束等价;

3) 软约束. 鼓励智能体的动作和状态满足约束, 可在硬约束的不等式右边加入非负项来实现, 即

$$c_t^j(s_t, a_t, s_{t+1}) \leqslant \epsilon_j, \tag{7}$$

并在目标函数中加入适当的惩罚项$l_\epsilon(\boldsymbol{\epsilon}) \geqslant 0$, 且$l_\epsilon(\boldsymbol{\epsilon}) = 0 \Leftrightarrow \boldsymbol{\epsilon} = \mathbf{0}$, 提高违反约束的成本, 向量$\boldsymbol{\epsilon}$作为优化变量.

在状态约束中, $c_t^j(s_t, a_t, s_{t+1})$是单步量, 描述了智能体每一时刻的约束条件. 这种单步约束的特性更为严格, 也更符合实际机器人应用场景的需求, 例如地面移动机器人系统需要确保在任何时刻都不会超出道路范围或撞到行人[4], 家政机器人则需要避免对家具进行损坏[30]. 在实际机器人系统中, 安全约束通常是混合类型的, 例如输入约束通常是硬约束, 而状态约束可能是软约束[24].

因此, 考虑满足约束的安全强化学习问题可以描述为在遵守安全约束条件的前提下, 寻求最大化期望累积回报的策略π^*[25], 即

$$\pi^* = \arg\max_{\pi} V^{\pi},$$
$$\text{s.t.}\ \ \text{安全约束(2), (3), (4), (5), (6), 和/或(7).} \tag{8}$$

这种框架为马尔可夫决策过程引入了一系列限制和约束. 在强化学习中, 解决安全约束问题面临两个关键挑战: 如何将约束条件融合到强化学习算法中, 以及如何有效地解决带有约束条件的强化学习问题.

2.2 考虑鲁棒性的RL问题

除了满足约束条件外, 安全性还要求智能体具备对系统不确定性和外界干扰等风险因素的鲁棒性. 为了增强鲁棒性, 安全强化学习问题通常可建模为鲁棒马尔可夫决策过程(RMDP)[27], 在存在不确定性或扰动时优化最坏情况下的值函数或成本函数, 可通过以下max-min优化问题来表示:

$$\pi^* = \arg\max_{\pi} \min_{f_{\rm d} \in D} V^{\pi}, \tag{9}$$

其中$f_{\rm d} \in D$为系统不确定性和扰动项. 鲁棒性安全强化学习的关键问题包括如何考虑不确定性和干扰因素, 以及如何有效地求解max-min优化问题.

3 安全RL算法与理论研究进展

本节从考虑安全约束和增强鲁棒性两类方法出发, 对当前安全强化学习算法和理论的研究进展进行概述. 这两类方法分别对应解决上节所述两种安全强化学习问题. 表1从解决问题的目标、方式、方法以及优缺点几个方面对本文综述的安全强化学习方法进行了直观的分析和比较.

3.1 考虑约束的安全RL方法

目前安全强化学习领域主流的方法是通过建立约束马尔可夫决策过程来引入安全约束信息. 针对约束马尔可夫决策过程, 求解方式主要有以下3种: 第1种通过修改学习目标, 将安全因素纳入目标函数中; 第2种通过修改学习过程, 限制智能体采取危险动作; 第3种通过学习约束信息, 在无法事先获得安全知识的情况下从经验中提取约束信息. 本小节由此3种方式出发, 介绍不同约束安全强化学习方法研究工作.

3.1.1 修改学习目标

常规强化学习方法的目标函数不考虑危险状态和动作造成的潜在损害, 可能带来安全问题. 为了解决

这一问题, 修改学习目标的方式将风险信息引入回报或成本函数, 继而进行约束优化问题的求解, 求解方式通常将原约束优化问题转化为无约束优化问题. 目前, 修改学习目标的方法主要包括拉格朗日法、置信域法和后向值函数法等, 此类方法通常适用于处理第2.1节中的轨迹约束问题.

表 1 安全强化学习方法分类及对比

Table 1 Classification and comparison of safe reinforcement learning methods

目标	方式	方法	优点	缺点
满足安全约束	修改学习目标	拉格朗日法[31–35]	使用方便、易于实现	拉格朗日乘子选取困难
		置信域法[36–40]	训练稳定、收敛性好	存在近似误差、计算代价较大
		后向值函数法[41]	解决状态级别约束、降低了计算代价	需要额外的安全层
	修改学习过程	李雅普诺夫函数法[42–44]	可以保证智能体闭环稳定性	李雅普诺夫函数构造困难、需要模型
		控制障碍函数法[45–48]	有效指导训练期间安全探索	障碍函数和权衡系数设计困难、需要模型
		外部干预[49–62]	加快学习过程、大大减小不安全因素	专家经验不易获得、人工成本高、学习过程不连续
	学习约束信息	安全评价器[63–66]	无需专家经验标记安全知识, 扩展性强	过滤准则依赖无模型的学习值函数, 训练不稳定
		贝叶斯推理[67–70]	天然地解决探索–利用问题	计算复杂度高
增强鲁棒性	鲁棒训练设计	对抗性博弈训练[71–77]	解决min-max问题思路直接、无需建模风险和不确定性	训练必需仿真环境、训练不稳定
		随机性泛化训练[78–81]	学习经验丰富、策略泛化能力强	合适的随机因子设计困难、需要较长的训练时间
	鲁棒性分析	可达性分析[82–88]	计算安全集得出更大的可行策略空间	需要探索安全边界, 难以保证训练时安全
		风险与不确定性建模[89–94]	直接通过推导和使用风险提高安全性, 直观有效	安全性依赖选择考虑的风险因素和对其的估计精度

1) 拉格朗日方法.

对于满足约束$g(x) \leqslant 0$的优化问题$\min\limits_{x} f(x)$, 拉格朗日方法引入了拉格朗日函数$\mathcal{L}(x,\lambda) = f(x)+\lambda g(x)$, 可以等价地在原始变量$x$和对偶变量$\lambda$上求解无约束优化问题$\max\limits_{\lambda\geqslant 0}\min\limits_{x}\mathcal{L}(x,\lambda)$, 原始–对偶求解常采用数值更新方法执行[31]. 该方法在解决约束强化学习问题时, 通过拉格朗日乘数法将安全约束添加到目标函数中, 在不违反或尽可能少违反安全约束的前提下寻找最优解[31–32]. 具体而言, 首先使用拉格朗日函数$\mathcal{L}(\pi,\lambda)$将约束优化问题转化为无约束优化问题, 然后利用强化学习方法来求解原始–对偶无约束优化问题, 即

$$\begin{cases}\mathcal{L}(\boldsymbol{\pi},\lambda) = V^{\pi} + \sum\limits_{j}\lambda_j\left(V_{c^j}^{\boldsymbol{\pi}} - d^j\right),\\ (\pi^*,\lambda^*) = \arg\max\limits_{\lambda\geqslant 0}\min\limits_{\pi}\mathcal{L}(\pi,\lambda).\end{cases}\tag{10}$$

使用拉格朗日方法, 文献[31]提出了一种用于风险约束强化学习的原始–对偶次梯度方法, 其中定义了基于条件风险值(conditional value-at-risk, CVaR)的约束优化问题. 在文献[31]的基础上, 文献[32]引入了对偶变量的离策略更新, 通过实证表明该更新方式具有更高的样本效率和更快的收敛速度. 然后, 文献[33]提出了一种约束执行器–评价器方法, 称为回报约束策略优化 (reward constrained policy optimization, RCPO), 首先通过拉格朗日乘子法将约束条件转化为惩罚信号, 然后在不同的时间尺度上更新执行器、评价器和拉格朗日乘子, 最后证明了方法的局部收敛性. 除此之外, 文献[34]对无模型的软执行器–评价器(soft actor-critic, SAC)进行了扩展, 在CMDP框架中定义行为偏好, 通过指定阈值和指示函数来规范行为. 该方法不需要人类先验知识或持续反馈, 利用拉格朗日方法自动权衡行为约束, 完成了基于目标的多约束任务. 文献[35]提出了一种保守的原始–对偶随机算法, 通过后悔分析证明了在不违反约束的情况下解决CMDP问题的可行性. 基于拉格朗日方法的安全强化学习方法在约束优化和强化学习之间建立了联系, 将约束优化问题简化为带有惩罚项的无约束优化问题. 然而, 这种方法仅适用于处理轨迹约束并渐进收敛, 在学习过程中不能保证系统的安全性, 并且对拉格朗日乘子的初始值和学习率的选择较为敏感.

2) 置信域方法.

置信域法在策略优化时与拉格朗日法有所不同, 置信域法显式表示安全约束, 在每次迭代过程中将策略投影到安全可行集内, 以控制策略的更新方向, 确保策略始终满足期望的约束. 文献[36]提出的约束策略优化(CPO)方法首次将置信域策略优化(trust region policy optimization, TRPO)[95]扩展到解决CMDP问题, 以限制单步策略更新不违反约束, 并在理论上保证了算法接近约束满足. 在实现中, CPO方法使用了拉格朗日方法来近似求解. 近年来, 基于CPO的算法在安全强化学习领域受到广泛关注, 并发展出多种解决约束MDP的算法[37–40]. 其中, 文献[37]提出了一种基于投影的约束策略优化 (projection-based CPO, PCPO)算法, 使用投影梯度下降将TRPO的中间策略投射回约束集. 另外, 文献[38]提出了一阶约束优化方法, 将更新的策略映射回参数化策略空间, 避免了使用Fisher逆矩阵, 提高了计算效率. 对于高维连续控制任务的仿真结果显示, 相较于复杂的二阶近似方法, 使用一阶近似能够获得更好的性能. 然后, 文献[39]提出了一种惩罚近端策略优化 (penalized proximal policy optimization, P3O)算法, 引入精确惩罚函数将原问题转化为无约束优化问题, 并采用一阶优化方法求解, 避免了逆矩阵计算困难的问题. 此外, 文献[40]提出了一种约束修正策略优化方法, 该方法基于原始问题进行求解, 采用自然策略梯度来更新策略, 在性能优化和约束满足之间交替更新策略, 不需要引入额外的对偶变量, 并分析了全局最优的收敛性. 置信域法可以在每次更新中近似保证满足轨迹约束, 在训练时有着较好的安全性, 但是在矩阵求逆时计算代价较大, 运用一阶或二阶近似的方法存在近似误差.

3) 后向值函数方法.

文献[41]提出了一种基于后向值函数 (backward value functions, BVFs)的约束马尔可夫过程求解方法, 解决了上述两种方法计算成本过高的问题. 类似于标准值函数V^{π}估计从当前时刻出发未来累积回报或成本的期望值, 后向值函数$V^{b,\pi}$估计了从初始状态到当前状态累积回报或成本的期望值. 然后, 可以将任意时刻t的轨迹级约束成本$V_{c_j}^{\pi}$分解为前向约束成本$V_{c_j}^{\pi}(s_t)$和后向约束成本$V_{c_j}^{\mathrm{b},\pi}(s_t)$的和, 再减去当前时刻的约束成本$c_j(s_t)$. 因此, 后向值函数方法可以将轨迹累积成本的约束转化为基于状态的约束. 此外, 文献[41]证明了转换后基于状态的约束函数值是原始约束函数值的上界, 即如果给定轨迹中每一步的状态约束得到满足, 那么就可以保证给定的CMDP约束得到满足. 后向值函数方法缓解了CPO算法需要回溯搜索过程和共轭梯度算法逼近Fisher信息矩阵时计算成本过高的问题. 在实际应用中, 可以利用从环境收集的样本通过时间差分(temporal difference, TD)[96]方法同时学习后向值函数和相应的约束成本, 从而缓解了离线评估的问题. 后向值函数方法提供了一种新的视角, 将轨迹约束分解为每个状态时刻的约束对求解约束马尔可夫决策过程. 但是在文献[41]中, 其策略需要添加文献[60]中提出的安全层.

3.1.2 修改学习过程

仅通过修改学习目标无法保证智能体在自由探索时不会面临危险和安全事故的问题. 因此, 为了避免在探索过程中出现意外情况, 往往需要对智能体的行动进行安全性评估, 并通过先验知识来限制其仅在安全区域内进行探索, 此方式可以叫做修改学习过程. 修改学习过程的方法包括李雅普诺夫函数法、控制障碍函数法以及外部干预法, 外部因素主要包括专家经验、人工介入和额外的安全模块等, 此类方法可以适用于处理第2.1节中的状态约束问题.

1) 李雅普诺夫函数方法.

李雅普诺夫函数在控制理论中广泛用于分析动态系统的稳定性, 是一种将全局属性转换为局部条件的重要工具[97]. 在强化学习中, 为了降低探索过程中的风险, 要求智能体能够从不安全状态恢复, 即具备控制理论中的渐近稳定性[25]. 在闭环系统的稳定性分析时, 给定某个策略$\pi(s)$, 系统可以表示为

$$s_{t+1} = f_{\pi}(s_t) = f(s_t, \pi(s_t)), \tag{11}$$

其中f表示系统的转移函数, 它是利普希茨连续的. 如果存在一个利普希茨连续的正定函数

$$L : S \to \mathbb{R}_{\geqslant 0}, \tag{12}$$

满足$L(\mathbf{0}) = 0$和$L(s) > 0, \forall s \neq \mathbf{0}$, 并且在闭环状态反馈下, 该函数将反馈状态映射到比原状态严格小的值上(即$\Delta L(x) = L(f_{\pi}(s)) - L(s) < 0$), 那么系统的状态将收敛于原点处的平衡态. 对于连续时间系统, 也有类似基于L导数的表达式. 文献[42]首次基于李雅普诺夫函数设计安全强化学习方法, 通过设计智能体在不同基准控制器之间的切换策略, 保证了系统的闭环稳定性. 为了降低离策略评估过程中的计算代价, 并更有效地处理轨迹约束(如式(2)), 文献[43–44]利用李雅普诺夫函数将其转化为逐步的基于状态的约束. 其中, 文献[43]的方法适用于离散动作空间问题, 而文献[44]则通过策略梯度方法将其扩展到连续动作空间问题. 李雅普诺夫函数方法的基本思想是将整体约束分解为一系列局部约束, 因此当状态空间无穷大时, 可行集会形成无穷维的约束集. 实现李雅普诺夫稳定性约束的策略更新在计算成本和方法适用性等方面仍具挑战性, 且该方法需要模型知识[25]. 此外, 此类方法要求系统的初始策略π_0满足约束条件, 但在某些复杂任务中难以找到满足约束的初始策略, 从而一定程度上限制了方法的应用范围.

2) 控制障碍函数方法.

障碍函数最初源于优化理论, 作为一类约束函数用于优化问题的求解[98]. 在控制领域中, 控制障碍函数(control barrier functions, CBFs)广泛应用于控制策略的安全性问题, 以确保系统的稳定性和安全性. 在安全强化学习中, 一些研究工作通过引入CBF约束限制智能体的动态行为, 从而保证智能体在执行任务时不会违反事先设定的安全约束条件. 具体而言, CBF是一种将状态空间映射到实数空间的连续可微函数$B_{\mathrm{c}}: S \to \mathbb{R}$, 安全集可表示为

$$\Omega_{\mathrm{safe}} = \{s : B_{\mathrm{c}}(s) \geqslant 0\}, \tag{13}$$

即CBF在可行集的内部是正定的. 文献[45]将无模型的强化学习方法与基于模型的CBF控制器以及高斯过程未知系统动力学建模相结合, 提出了一种控制器架构确保学习过程中的安全性, 并实现了端到端的安全强化学习. 文献[46]通过障碍函数对系统施加状态约束, 并将原始问题转换为无约束优化问题, 然后提出了静态和动态两种间歇反馈强化学习算法, 利用执行器–评价器(actor-critic, AC)框架实现了安全策略的在线学习. 文献[47]将控制障碍函数增广到离策略强化学习算法的成本函数中, 并为CBF设计了权衡因子以平衡安全性与最优性. 该方法实现了在安全集内状态的平稳过渡, 并在车道保持实验中进行了验证. 文献[48]在约束边界上定义了一种广义控制障碍函数(generalized CBF), 针对约束强化学习提出了一种基于模型的可行性增强技术. 利用模型信息, 该方法能够在不违反实际安全约束的情况下进行策略优化, 提高了样本效率, 并通过自适应系数机制解决了约束策略梯度求解的不可行性问题. 控制障碍函数方法在安全强化学习中提供了有效的思路, 将安全集约束融入到强化学习中并指导训练过程中的安全探索. 然而, 在实际应用中, 障碍函数仍需要人工设计和选择, 并且权衡系数的调节较为困难.

3) 外部干预方法.

外部因素, 如专家指导[49–52]、人工介入[53–54]以及额外的安全过滤模块[55–59]等, 可以指导智能体的学习过程或在发现不安全动作时对智能体系统进行中断或修正. 在文献[50]中, 受课程教学方法的启发, 智能体在自动教师(automatic instructor)的监督下进行学习, 并引入监视器, 在智能体做出危险动作时进行重置操作, 避免了智能体在学习过程中违反约束. 针对智能体在训练过程中的安全性问题, 文献[49]采用基于模仿学习[99]的人工介入方式保证智能体的安全运行. 另一方面, 文献[55]提出了一种基于屏障(shielding)的安全学习系统, 该屏障模块监视来自执行器的动作, 并在动作导致违反约束时进行纠正. 这种使用安全层[60]或安全过滤器[56]等模块的方法通常对原始动作的纠正方式如下:

$$\begin{aligned} &a_{\mathrm{safe},t} = \arg\min_{a_t \in A} \|a_t - a_{\mathrm{learn},t}\|_2^2, \\ &\text{s.t.} \quad s_{t+1} \in \Omega_{\mathrm{safe}}. \end{aligned} \tag{14}$$

类似地, 文献[59]将屏障策略和优势函数干预机制相结合, 借助屏障策略替换原策略中的不安全动作, 并在屏障策略触发时, 通过优势函数干预机制对原策略进行惩罚. 基于屏障的方法简单有效地保证了安全, 但是其安全性往往受限于安全层模块的性能. 此外, 一些研究通过引入其他的安全模块来确保智能体的安全性. 在文献[61]中, 提出了一种结合强化学习和蒙特卡洛树搜索的算法, 主要包括风险状态估计和安全策略搜索两个模块, 如果计算得到未来状态存在风险, 就会通过增加风险动作的额外惩罚保证更安全的探索. 文献[62]提出一种双层控制器架构, 其中上层为元感知层, 用于关注系统的安全性和性能要求, 下层则利用强化学习算法结合上层信息来学习最优控制策略, 由于元感知层的存在, 该架构可以在不同情况和不同任务中进行安全强化学习. 综上所述, 专家经验的引入可以加速学习过程, 而人工介入和额外的安全过滤模块可以显著减少不安全因素. 然而, 此类方法也存在一些缺点, 首先, 获取专家经验并将其应用于学习过程并非易事; 其次, 人工介入会增加人力成本; 此外, 频繁中断可能导致学习过程的不连续.

3.1.3　学习约束信息

在某些情况下, 环境中的安全知识或约束信息无法提前获得, 这给人工设定安全约束带来了一定的困难. 因此, 一类研究工作聚焦对安全知识进行在线学习, 这类方法主要包括安全评价器方法和贝叶斯推理, 旨在没有先验知识的情况下, 通过学习或推理环境中的约束和风险知识, 达到提高安全性的目的.

1) 安全评价器方法.

安全评价器方法通过学习一个安全评价器(safety critic) Q^{π}_{safe}来评价当前状态和动作的安全性. 文献[63]提出了一种安全Q函数强化学习(safe Q-functions for RL, SQRL)算法, 利用安全动作–状态值函数Q^{π}_{safe}从环境中学习稀疏的安全特征, 并可以迁移到相似的任务中. 在SQRL中, Q^{π}_{safe}用来预测未来轨迹的不安全概率, 用于对危险动作进行过滤. 该方法在实现时, 首先, 在仿真环境中学习安全值函数和策略, 然后, 将策略迁移到实际系统中. 基于SQRL, 文献[64]提出了一种可恢复强化学习方法, 通过学习恢复策略π_{rec}作为在原策略经过危险动作过滤后的替代方案. 文献[65]在保守Q学习[100]算法的基础上, 提出了一种保守安全评价器方法, 用于学习环境的安全特征并过滤危险动作. 该方法从理论上描述了安全性和策略改进之间的权衡, 并证明了训练过程中安全约束的概率满

足. 文献[66]提出了一种最坏情况软执行器–评价器(worst-case soft actor critic)算法, 研究了对安全值函数的两种估计方法, 分别为高斯近似和分位数回归, 并分析得出在复杂安全约束环境下, 分位数回归方法能够取得更好的效果. 安全评价器方法无需专家经验标记安全知识, 具有扩展性强的优点, 但其对不安全动作的过滤准则依赖于无模型学习的值函数, 难以保证学习期间的安全性.

2) 贝叶斯推理方法.

贝叶斯优化(Bayesian optimization)是一种用于优化高成本目标函数的方法, 依靠概率和贝叶斯推理来推演优化过程中的不确定量, 已经在不同领域取得了成功的应用[101]. 一种基于贝叶斯优化的安全学习算法是SafeOpt(safe Bayesian optimization)[102], 它使用高斯过程(Gaussian processes, GP)对安全集和目标函数进行建模和推断, 并在优化过程中对满足所有安全约束的参数进行概率评估. SafeOpt的扩展算法还包括多约束安全贝叶斯优化(SafeOpt multiple constraints, SafeOpt-MC)[103], 阶段安全贝叶斯优化(stage-wise safe Bayesian optimization, StageOpt)[104]以及全局最优安全机器人学习(globally optimal safe robot learning, GoSafe)[105]等. 进而, 一些研究工作将贝叶斯优化的思想用于安全强化学习, 以解决强化学习的探索和利用权衡问题[67], 这类方法通常使用高斯过程对未知目标函数进行建模和学习. 在基于贝叶斯优化的安全学习方法启发下, SafeMDP方法[68]将SafeOpt的思想引入到马尔可夫决策过程中, 在未知成本函数的情况下进行安全探索, 并可满足单步安全约束. SafeExpOpt-MDP方法[69]将安全函数和MDP中的值函数作为独立的未知函数使用高斯过程进行学习, 以在优化智能体性能的同时满足安全约束. 此外, 综述文献[70]详细阐述了安全学习、贝叶斯优化和高斯过程建模之间的联系. 贝叶斯强化学习框架提供了一种解决探索和利用问题的原则性方法, 其主要优点是在选定的参数表示中考虑了状态的全部信息, 包括安全信息. 然而, 目前贝叶斯强化学习应用的主要挑战在于选择合适的信息表示方法, 并且定义基于信息的状态往往要比原始状态更为复杂.

3.2 增强鲁棒性的安全RL方法

增强鲁棒性的安全强化学习旨在实现智能体策略在面对不确定性或干扰等风险因素时的安全学习和部署, 通常可以建模为鲁棒马尔可夫决策过程(9). 本节将介绍两种增强强化学习策略鲁棒性的方式, 分别为鲁棒训练设计和鲁棒性分析.

3.2.1 鲁棒训练设计

鲁棒训练设计旨在改进智能体的训练过程, 以获得对干扰或不确定性鲁棒的策略, 主要包括对抗性博弈训练和随机性泛化训练. 对抗性博弈训练使用博弈论的思想解决max-min优化问题, 在此类方法中, 智能体在训练过程中与其他智能体或环境中的对手进行竞争, 并且尝试找到最佳策略来应对可能的对手行为, 通过与对手交互, 智能体可以学习适应不同情境下的干扰. 随机性泛化训练增加系统的不确定性或引入扰动进行学习, 训练过程引入一定程度的随机性, 使智能体能够学习应对不确定性.

1) 对抗性博弈训练方法.

文献[71]将强化学习与对抗性学习[106]相结合, 提出了一种鲁棒对抗性强化学习(robust adversarial RL, RARL)方法. 该方法将鲁棒强化学习问题设置为两方零和博弈的马尔可夫过程, 其中一方智能体(主角)学习策略π控制系统; 另一方智能体(对手)学习策略破坏系统. 两个智能体以交替的方式进行学习和更新, 相互影响, 逐步提高主角策略对干扰的鲁棒性. 文献[72]在RARL的基础上重点关注智能体对风险的处理, 其中主角避免风险, 而对手寻求风险, 其中风险定义为值函数预测的方差, 并且该方法结合了深度Q网络(deep Q-networks, DQN)[73]进行学习. 文献[74]将RARL中的对手扩展为对手群, 以降低算法对单一对手的依赖, 进一步提高主角策略的鲁棒性. 此外, 文献[75]提出一种鲁棒强化学习方法, 在执行过程中计算状态–动作值的下界, 以在对手或不确定性造成最坏情况时选择最优动作. 然后, 文献[76]提出了部分监督强化学习框架(partially- supervised RL, PSRL), 将对抗性强化学习置于确定性的部分可观马尔可夫决策过程中, 以达到在安全约束下最大化累积奖励的目的. 另外, 文献[77]设计了一种执行器–干扰器–评价器结构, 其中干扰器的目标是制造最恶劣的干扰, 而执行器的目标是产生最佳的控制输入. 研究者观察到, 通过该方法学得的策略和值函数与线性系统H_∞控制理论分析推导的策略和值函数一致. 对抗性博弈训练方法直接考虑鲁棒强化学习的min-max问题, 无需对风险和不确定性进行建模, 但由于在博弈过程中对手寻求最坏情况可能导致危险发生, 因此, 该类方法的训练过程依赖仿真环境.

2) 随机性泛化训练方法.

与考虑最坏情况干扰的对抗性博弈训练不同, 随机性泛化训练方法在训练过程中增加系统的不确定性和扰动, 这些不确定性通常在预定义的范围内, 并可以表示为不确定性集合D. 然后可以使用标准强化学习算法进行训练, 主要目的是使智能体能够学习应对不确定性, 提高策略的泛化性能. 文献[78]在具有随机属性的仿真环境中, 引导四旋翼飞行器进行了基于视觉的飞行学习, 所学得的鲁棒策略可以迁移到真实世界环境中并具备避碰性能. 类似地, 文献[79]中的工作也在仿真环境中进行学习, 通过域随机化(domain

randomization)技术产生丰富的模拟数据. 文献[80]提出了一种集成策略优化(ensemble policy optimization, EPOpt)算法, 利用经验数据和近似贝叶斯方法, 对仿真环境中随机参数的概率分布进行自适应调整, 以产生高质量的数据并更好地学习鲁棒策略. 在文献[81]中, 针对域随机化可能会导致次优和高方差策略的问题, 提出了一种主动域随机化算法, 通过对抗性博弈训练寻找当前策略较少探索或利用的随机化区域, 以更高效地学习近似最优的鲁棒策略. 随机性泛化训练旨在通过丰富的经验数据提高策略的泛化性能, 然而, 该方法在训练过程中需要选择适当的随机因子, 并且相较于常规方法需要更长的训练时间.

3.2.2　鲁棒性分析

与鲁棒训练设计方法不同, 鲁棒性分析的方法主要通过对不确定性或者潜在的风险进行分析、建模和估计, 在学习策略时充分考虑不确定性, 以提高对不确定性或风险的鲁棒性, 主要包括可达性分析方法和风险与不确定性建模方法. 可达性分析方法提供了一种安全集分析方法, 用于对系统的状态空间进行分析, 以确定系统可以达到的安全状态集合. 这种方法通过限制策略的探索过程, 确保系统在面对不确定性时能够保持在安全状态. 风险与不确定性建模方法通常基于统计模型或机器学习技术, 对环境中的风险因素或不确定性进行建模, 并利用这些模型知识来指导策略的学习过程.

1) 可达性分析方法.

哈密尔顿–雅可比可达性分析(Hamilton-Jacobi reachability analysis, HJ-RA)方法源于鲁棒最优控制理论, 是一种通用严谨的可行集分析方法[107]. 该方法通过以下步骤进行: 首先, 设函数$l: S \to \mathbb{R}$, 满足$l(s) \geqslant 0$, 当且仅当s为安全状态. 然后, 定义最优安全值函数$V_l(s)$如下:

$$V_l(s) := \max_{\pi} \inf_{t \geqslant 0} l(s|s_0 = s), \tag{15}$$

表示从状态s出发, 使用策略π与环境交互的轨迹中函数$l(s)$的最小值. 安全集Ω_{safe}的定义为$\{s: V_l(s) \geqslant 0\}$, 表示函数$l(s)$的最小值大于零, 即使用策略$\pi$与环境交互的轨迹是安全的[82]. 根据此定义, 可以计算最优安全策略π^*_{safe}, 其在最快上升方向上最大化V_l, 以将系统导向安全集Ω_{safe}. HJ可达性分析方法考虑现实中的安全问题通常为伴随着时间推移而出现的最坏情况. 文献[82]首次将HJ可达性分析引入无模型强化学习算法中, 以对安全约束进行严格形式化. 然后, 文献[83]提出了一种基于HJ可达性方法的通用安全框架, 该框架可以与任意学习算法结合使用, 并使用高斯过程对系统动力学进行近似, 只在安全性能需要时进行策略干预. 文献[84]结合了HJ可达性分析和控制障碍函数, 以实现更鲁棒的控制策略, 同时避免了人工设计控制障碍函数, 将可达性分析方法与控制障碍函数相结合以获得更好的性能. 最近的拓展研究工作[85]改进了HJ可达性分析方法在高维系统上的适应性, 并在十维四旋翼无人机跟踪控制问题上进行了验证. 文献[86]首先对机器人系统的黑盒模型进行数据驱动的可达性分析; 然后, 使用在线训练的神经网络预测未来轨迹; 最后, 通过可达集与障碍物之间的碰撞关系来纠正不安全行为. 此外, 文献[87]针对基于HJ可达性分析的安全强化学习算法的最优性问题, 提出了一种可达性约束强化学习(reachability constrained RL, RCRL), 通过将可达性约束引入约束优化中, 在可行集上学习接近最优且持续安全的策略, 并证明了该策略在最大可行集上满足约束. 随后, 文献[88]将RCRL拓展到基于模型的强化学习方法中. HJ可达性分析方法提供了一种适用的安全集分析方法, 可以计算出更大的可行策略空间, 并在与强化学习的结合中展现出良好的效果. 然而, 这种方法在学习过程中需要探索安全边界, 因此难以保证训练时的安全性.

2) 风险与不确定性建模方法.

该方法通过在强化学习过程中推导和使用风险或不确定性, 以训练出更保守的策略来提高学习的安全性. 在文献[89–90]中, 风险被定义为机器人系统在执行任务时发生碰撞的概率. 这些研究使用神经网络对碰撞模型进行学习, 得到了给定当前状态和未来动作序列的碰撞概率分布, 然后在学习过程中考虑碰撞模型以选择执行的动作. 继而, 文献[91]提出了一种谨慎适应安全强化学习(cautious adaptation for safe RL, CARL)方法, 利用模型集成和概率网络评估系统动力学中的不确定性[108], 并提出了避免低回报轨迹与避免灾难性碰撞两个风险概念. 文献[92]提出了一种离线风险规避的执行器–评价器方法(offline risk-averse actor-critic, O-RAAC), 使用条件风险值评估风险, 能够在完全离线的环境中学习风险规避策略, 并且该方法无需使用模型. 此外, 文献[93]提出了均值–方差策略迭代(mean-variance policy iteration, MVPI)方法, 在MDP中使用均值和方差对风险进行度量, 并以双延迟深度确定性策略梯度算法(twin delayed deep deterministic policy gradient algorithm, TD3)为例, 在仿真环境中实现了风险规避强化学习. 类似地, 文献[94]选择了均值–半方差(mean-semivariance, MSV)作为风险度量, 即随机变量在其均值下的负偏差. 本节讨论的风险与不确定性建模方法和第3.1.3节中学习约束信息的方式均关注学习环境中的风险知识, 但其重点略有不同. 不同的是前者关注学习或部署过程中如何考虑最坏风险或者不确定性, 倾向于学习保守的策略, 而后者关注解决约束优化问题, 通过学习和应用先验的安全知识来限制策略的选择空间. 风险与不确定性

建模方法直接有效地通过推导和使用风险估计来提高系统的安全性, 但是其性能依赖于选择考虑的风险因素和估计精度.

4 安全RL在机器人系统中的应用研究进展

安全强化学习方法已经在机器人领域获得了广泛的关注, 其目标是确保强化学习在机器人系统上的训练和部署过程中能够保持安全和合理的行为. 在最近的研究工作中, 安全强化学习已经应用于设计地面移动机器人、无人飞行器和工业机器人等系统的决策、规划和控制算法, 以确保它们在面对不可预测的情况时采取正确的行动, 例如避免碰撞和遵守交通规则等, 为机器人系统的自主智能操作提供了一种有效的途径, 并有望使其更加安全、高效、准确地完成任务. 下面从安全强化学习在地面移动机器人、无人飞行器以及其他机器人系统中的应用研究情况展开详细阐述.

4.1 安全RL在地面移动机器人系统中的应用

地面移动机器人系统需要在复杂和不可预测的交通环境下实现精确且安全的行驶. 文献[109]将深度确定性策略梯度(deep deterministic policy gradient, DDPG)[110]和安全控制算法相结合, 并使用人工势场法进行机器人路径规划. 该方法首先在稳定的环境中使用DDPG学习驾驶策略, 然后结合策略网络和安全控制避免碰撞, 并表明在大多数情况下, 深度强化学习与安全控制相结合能够取得良好的性能. 文献[111]针对理想化模型对现实世界随机性和高维性的适应性问题, 研究了如何使用模型预测来约束探索过程, 并重点关注了地面移动机器人系统在十字路口的行为. 文献[112]提出了一种增强策略概率保证的方法, 在训练之前推导出一个探索策略, 使智能体在动作空间中选择的动作满足线性时序逻辑(linear temporal logic, LTL)的期望概率规范, 以缩小搜索空间并简化奖励设计, 该研究在涉及多个交通参与者的交叉路口情景案例中进行了应用实验. 文献[113]在强化学习中加入人类决策模型, 采用后悔理论(regret theory)来描述和建模人类驾驶员的换道行为, 然后将人类驾驶员决策纳入安全条例, 以控制地面移动机器人系统进行安全有效并符合人类期望的操作. 文献[114]针对地面移动机器人系统的规划任务提出了一种层次结构, 引入了策略导向的轨迹规划器, 上层深度强化学习算法输出策略, 对下层轨迹规划器进行导向和建议. 这种方法将交通规则和策略转化为正式的规范, 上层深度强化学习算法处理长期的不确定性任务, 而下层轨迹规划器负责保证安全性. 文献[69]针对车道保持和多车决策任务的安全约束问题, 提出了一种并行约束策略优化(parallel constrained policy optimization, PCPO)方法, 将目前常见的执行器–评价器两学习器结构扩展为三学习器结构, 使用3个神经网络分别逼近策略函数、值函数和风险函数, 在搜索最优策略过程中同时考虑性能和风险. 在文献[69]工作的基础上, 文献[115]提出了一种基于控制障碍函数安全层的决策方法[116], 并通过遵循交通规则确保高速公路场景下自动驾驶运动规划的安全性和合法性. 文献[117]提出了一种多时间尺度约束DQN算法, 该算法引入了基于截断值函数的近似多步约束公式, 直接在Q函数更新中限制动作空间, 能够在自动驾驶仿真平台[118]中学习策略, 并迁移到真实数据集[119], 验证了该算法的有效性. 文献[120]使用变分自编码器[121]和DDPG构建了一个无模型强化学习框架, 并加入逻辑规则, 实现了真实世界中的车辆安全驾驶实验. 文献[122]提出了一种多步策略评估机制, 用于预测时变CBF安全约束下的策略风险. 该机制能够引导地面移动机器人系统进行策略安全更新, 并能够在理论上证明系统的稳定性和鲁棒性.

4.2 安全RL在无人飞行器系统中的应用

与地面移动机器人系统相比, 无人飞行器(unmanned aerial vehicles, UAVs) 具有更高维度的状态动作空间, 使得其安全学习任务更具挑战. 文献[123]为了应对微型飞行器控制的不可预测性和安全性问题, 并解决训练样本需求量大的困扰, 运用分层强化学习在未知环境下对安全性问题进行了扩展, 并在无人机仿真平台中进行训练. 文献[124]提出了一种基于两阶段强化学习的多无人机避碰方法, 在第1阶段使用监督学习方法训练策略使智能体学习共同的避碰规则, 在第2 阶段使用策略梯度方法进一步优化上述避碰策略, 目标是仅通过局部噪声观测完成无人飞行器无碰撞轨迹规划任务. 另外, 文献[79]展示了四旋翼无人机竞速的案例, 面对小型自主无人机环境动态变化、状态估计不可靠和计算资源限制等挑战, 实现了基于视觉的自主无人机竞速. 该工作通过域随机化方法在仿真环境中产生了丰富的模拟数据, 实现了增强鲁棒性的安全强化学习, 并将算法部署在竞速无人机的仿真和物理平台, 使其能够完成移动门高速穿越任务. 文献[125]提出了一种介入辅助式强化学习框架, 通过人工介入智能体–环境交互的方式改进策略, 并以无人机作为测试平台. 实验结果表明, 该算法能够减少人为干预, 提高自主导航性能, 并确保安全性. 在机器学习中, 选择合适的调优参数是一个普遍存在的难题, 显著影响算法的性能. 针对这一问题, 文献[103]利用了贝叶斯优化方法和SafeOpt思想, 将多个安全约束与目标分开考虑, 提出了一种名为SafeOpt-MC的算法. 该算法首先给定一组初始安全参数, 在优化过程中只对满足所有安全约束条件的参数进行高概率评估, 并可利用上下文变量将安全知识转移到新任务中. 在四旋翼飞行器系统中的实验表明, 该算法能够快

速、自动且安全地调整并优化参数. 文献[126]提出了一种强化学习的安全探索方法, 通过预定义风险函数和基准策略来探索安全空间, 并在直升机控制和自动泊车等任务中对提出方法进行了验证. 此外, 文献[127]提出了一种基于策略优先级的强化学习层次结构, 将策略划分为具有不同选择优先级的子策略, 从而降低了动作集的复杂性, 并通过智能体之间的迁移学习来初始化学习参数, 加速了最优策略的初始探索, 该方法在无人机的抗干扰通信中得到了应用.

4.3　安全RL在其他机器人系统中的应用

除上述两种典型的机器人系统外, 安全强化学习算法也在仿生机器人、工业机器人和机械臂等系统中得到了广泛的应用研究. 机器人系统通常需要进行导航、行走、抓取等基础任务, 并对人类工作进行辅助. 无论在生活应用还是工业控制场景中, 保证机器人系统的安全性都至关重要. 文献[128]针对6自由度工业机械手开发了一种完全可微的安全层, 该层将不安全的动作作为输入, 并输出满足约束的相近动作, 以确保机器人系统只能采取满足安全约束的动作. 文献[129]使用安全强化学习来改善类人机器人的行走行为: 首先, 假设存在一个安全的基准策略; 然后, 通过概率重用来学习更好的策略; 最后, 在人形机器人系统[130]中进行了实验验证, 结果表明该方法提高了学习速度并减少了学习过程中的跌倒次数. 文献[131]针对四足机器人的安全运动控制问题, 提出了一种安全强化学习策略框架, 在安全恢复策略和原策略之间进行切换, 其中安全恢复策略防止机器人进入不安全状态, 降低行走过程中的损伤风险, 并且尽量减少对学习过程的干预. 该方法在四足机器人仿真和实际系统中进行了验证, 可以执行走路、猫步、两腿平衡和踱步等任务. 文献[132]提出了一种接触式安全强化学习框架, 展示了安全强化学习在解决复杂接触式机器人操纵任务方面的潜力. 该方法考虑任务空间和关节空间的安全性, 当强化学习策略引起机器人手臂与环境之间的意外碰撞时, 能够立即检测碰撞并保持接触力. 在机器人擦拭任务上的实物实验表明, 该方法能够在任务中保持较小的接触力, 并且可以防止意外碰撞情况对任务的干扰.

5　未来研究展望

目前, 安全强化学习及其在机器人系统中的应用研究已取得了一些成果, 但仍面临着一些问题和挑战. 在本节中, 针对该领域研究的挑战和未来研究方向, 主要总结以下几点进行探讨:

1) 系统安全性和性能之间的权衡. 强化学习方法需要在未知的环境中进行试错学习, 然而过于激进的探索可能会导致机器人系统做出错误的决策, 从而造成意外危险, 过于保守又会对性能的提升产生不利影响. 因此, 强化学习中探索与利用的权衡至关重要. 在进行探索时, 必须考虑到安全性因素, 例如, 采用基于模型的强化学习方法, 即利用已有的环境模型指导机器人系统的探索, 使其更加高效和安全, 或者采用混合智能方法, 将传统的规则系统或人类经验和强化学习相结合, 实现对机器人系统的指导和控制, 以保证系统的安全性和性能.

2) 仿真和实际系统之间的策略迁移, 目前, 安全强化学习仿真平台如Safety Gym[133], Safe-Control-Gym[134], SafeRL-Kit[135], D4RL[136] 和NeoRL[137]等已经在模拟复杂物理机器人系统的任务中取得了积极成果. 在仿真环境中进行训练的优点为试错成本低, 安全性要求弱并且数据更易获得. 然而, 在实际系统中存在各种不确定性和噪声, 而且系统动态变化更加复杂, 仿真平台仍无法完全模拟实际系统. 此外, 物理世界的机器人系统在运行过程中需要更加严格的安全性能保证. 因此, 如何将在仿真环境中学习到的策略安全有效地迁移到实际系统中是未来研究的重要方向.

3) 在线学习中的安全性保证. 尽管目前的安全强化学习方法已经取得了许多进展, 但真实世界中在线强化学习的安全性理论和算法还研究不够充分. 此外, 由于真实世界中的运行成本较高, 智能体可能只能获得少量的数据(相较于仿真环境), 因此, 在线安全强化学习对样本利用效率也提出了很高的要求. 因此, 如何在现实世界中实现安全高效的在线学习, 并保证策略函数和值函数的收敛性, 仍然是一个具有挑战性的问题.

4) 机器人系统与人类的安全交互问题. 在机器人系统与人类交互的任务中, 由于人类行为的不可预测性, 机器人系统需要实时学习人类安全约束信息以及风格偏好, 以及时调整策略. 此类机器人系统需要保证在与人类互动的过程中安全可靠, 并能够通过不断学习和自适应来优化自身的行为. 因此, 在与人类的交互过程中, 如何使机器人系统具备强大的安全学习和自适应能力是一个重要问题.

6　结论

本文对近年来安全强化学习算法和理论进展及其在机器人系统中的应用进行了全面的综述和分析. 首先对安全强化学习问题进行了形式化建模与分类, 分为考虑安全约束的强化学习问题与考虑鲁棒性的强化学习问题. 然后分别阐述了针对不同类型问题的代表性研究工作, 包括考虑满足安全约束的修改学习目标、修改学习过程和学习约束信息, 以及考虑增强鲁棒性的鲁棒训练设计和鲁棒性分析等. 针对未来的研究, 重点从系统安全性和性能之间的权衡、在线学习中的安全性保证等几个方面进行了探讨.

参考文献:

[1] MURPHY R R. *Introduction to AI Robotics*. Cambridge, MA: MIT Press, 2019.

[2] DORIGO M, THERAULAZ G, TRIANNI V. Reflections on the future of swarm robotics. *Science Robotics*, 2020, 5(49): eabe4385.

[3] GOEL R, GUPTA P. Robotics and Industry 4.0. *A Roadmap to Industry 4.0: Smart Production, Sharp Business and Sustainable Development*. Cham: Springer, 2020, 157 – 169.

[4] KIRAN B R, SOBH I, TALPAERT V, et al. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 2022, 23(6): 4909 – 4926.

[5] KIM J, KIM S, JU C, et al. Unmanned aerial vehicles in agriculture: A review of perspective of platform, control, and applications. *IEEE Access*, 2019, 7: 105100 – 105115.

[6] HAGELE M, NILSSON K, PIRES J N, et al. *Industrial Robotics*, Cham: Springer International Publishing, 2016: 1385 – 1422.

[7] BRUNKE L, GREEFF M, HALL A W, et al. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 2022, 5(1): 411 – 444.

[8] XU X, YANG H, LIAN C, et al. Self-learning control using dual heuristic programming with global laplacian eigenmaps. *IEEE Transactions on Industrial Electronics*, 2017, 64(12): 9517 – 9526.

[9] ZHANG X, LIU J, XU X, et al. Robust learning-based predictive control for discrete-time nonlinear systems with unknown dynamics and state constraints. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2022, 52(12): 7314 – 7327.

[10] XU X, ZUO L, HUANG Z. Reinforcement learning algorithms with function approximation: Recent advances and applications. *Information Sciences*, 2014, 261: 1 – 31.

[11] FRANÇOIS-LAVET V, HENDERSON P, ISLAM R, et al. An introduction to deep reinforcement learning. *Foundations and Trends® in Machine Learning*, 2018, 11(3/4): 219 – 354.

[12] LECUN Y, BENGIO Y, HINTON G. Deep learning. *Nature*, 2015, 521(7553): 436 – 444.

[13] VINYALS O, EWALDS T, BARTUNOV S, et al. Starcraft ii: A new challenge for reinforcement learning. *ArXiv Preprint*, 2017: arXiv: 1708.04782.

[14] AFSAR M M, CRUMP T, FAR B. Reinforcement learning based recommender systems: A survey. *ACM Computing Surveys*, 2022, 55(7): 145.

[15] KAUSHIK P, SHARMA A R. Literature survey of statistical, deep and reinforcement learning in natural language processing. *2017 International Conference on Computing, Communication and Automation)*. Noida, India: IEEE, 2017: 350 – 354.

[16] MISRA S, DEB P K, KOPPALA N, et al. S-nav: Safety-aware iot navigation tool for avoiding covid-19 hotspots. *IEEE Internet of Things Journal*, 2021, 8(8): 6975 – 6982.

[17] JI Ying, WANG Jianhui. 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).

[18] LIU J, HUANG Z, XU X, et al. Multi-kernel online reinforcement learning for path tracking control of intelligent vehicles. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2021, 51(11): 6962 – 6975.

[19] KOBER J, PETERS J. Policy search for motor primitives in robotics. *Machine Learning*, 2011, 84(1/ 2): 171 – 203.

[20] MERIÇLI Ç, VELOSO M. Biped walk learning through playback and corrective demonstration. *AAAI Conference on Artificial Intelligence*. Atlanta, GA: AAAI, 2010, 24(1): 1594 – 1599.

[21] WU X, LIU S, ZHANG T, et al. Motion control for biped robot via ddpg-based deep reinforcement learning. *WRC Symposium on Advanced Robotics and Automation*. Beijing, China: IEEE, 2018: 40 – 45.

[22] ALTMAN E. *Constrained Markov Decision Processes*. volume 7. CRC Press, 1999.

[23] GARCIA J, FERN F. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 2015, 16(1): 1437 – 1480.

[24] GU S, YANG L, DU Y, et al. A review of safe reinforcement learning: Methods, theory and applications. *ArXiv Preprint*, 2022: arXiv: 2205.10330.

[25] WANG Xuesong, WANG Rongrong, CHENG Yuhu. Safe reinforcement learning: A survey. *Acta Automatica Sinica*, 2023, 49(9): 1813 – 1835.
(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).

[26] SUTTON R S, BARTO A G. *Reinforcement Learning: An Introduction*. London, England: MIT Press, 2018.

[27] NILIM A, EL GHAOUI L. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 2005, 53(5): 780 – 798.

[28] GEIBEL P. Reinforcement learning for mdps with constraints. *The 17th European Conference on Machine Learning*. Berlin, Germany: Springer-Verlag, 2006: 646 – 653.

[29] REGAN K, BOUTILIER C. Regret-based reward elicitation for Markov decision processes. *ArXiv Preprint*, 2012: arXiv:1205.2619.

[30] CRUZ F, TWIEFEL J, MAGG S, et al. Interactive reinforcement learning through speech guidance in a domestic scenario. *International Joint Conference on Neural Networks*. Killarney, Ireland: IEEE, 2015: 1 – 8.

[31] CHOW Y, GHAVAMZADEH M, JANSON L, et al. Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research*, 2017, 18(1): 6070 – 6120.

[32] LIANG Q, QUE F, MODIANO E. Accelerated primal-dual policy optimization for safe reinforcement learning. *ArXiv Preprint*, 2018: arXiv: 1802.06480.

[33] TESSLER C, MANKOWITZ D J, MANNOR S. Reward constrained policy optimization. *ArXiv Preprint*, 2018: arXiv: 1805.11074.

[34] ROY J, GIRGIS R, ROMOFF J, et al. Direct behavior specification via constrained reinforcement learning. *ArXiv Preprint*, 2021: arXiv: 2112.12228.

[35] BAI Q, BEDI A S, AGARWAL M, et al. Achieving zero constraint violation for constrained reinforcement learning via primal-dual approach. *AAAI Conference on Artificial Intelligence*. Electr Network: AAAI, 2022, 36(4): 3682 – 3689.

[36] ACHIAM J, HELD D, TAMAR A, et al. Constrained policy optimization. *ArXiv Preprint*, 2017: arXiv: 1705.10528.

[37] YANG T Y, ROSCA J, NARASIMHAN K, et al. Projection-based constrained policy optimization. *ArXiv Preprint*, 2020: arXiv: 2010.03152.

[38] ZHANG Y, VUONG Q, ROSS K W. First order constrained optimization in policy space. *ArXiv Preprint*, 2020: arXiv: 2002.06506.

[39] ZHANG L, SHEN L, YANG L, et al. Penalized proximal policy optimization for safe reinforcement learning. *ArXiv Preprint*, 2022: arXiv: 2205.11814.

[40] XU T, LIANG Y, LAN G. Crpo: A new approach for safe reinforcement learning with convergence guarantee. *International Conference on Machine Learning*. Electr Network: PMLR, 2021: 11480 – 11491.

[41] SATIJA H, AMORTILA P, PINEAU J. Constrained Markov decision processes via backward value functions. *ArXiv Preprint*, 2020: arXiv: 2008.11811.

[42] PERKINS T J, BARTO A G. Lyapunov design for safe reinforcement learning. *Journal of Machine Learning Research*, 2003, 3: 803 – 832.

[43] CHOW Y, NACHUM O, DUENEZ-GUZMAN E, et al. A Lyapunov-based approach to safe reinforcement learning. *ArXiv Preprint*, 2018: arXiv: 1805.07708.

[44] CHOW Y, NACHUM O, FAUST A, et al. Lyapunov-based safe policy optimization for continuous control. *ArXiv Preprint*, 2019: arXiv:1901.10031.

[45] CHENG R, OROSZ G, MURRAY R M, et al. End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. *AAAI Conference on Artificial Intelligence*. Honolulu, HI: AAAI, 2019, 33(1): 3387 – 3395.

[46] YANG Y, VAMVOUDAKIS K G, MODARES H, et al. Safe intermittent reinforcement learning with static and dynamic event generators. *IEEE Transactions on Neural Networks and Learning Systems*, 2020, 31(12): 5441 – 5455.

[47] 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.

[48] MA H, CHEN J, EBEN S, et al. Model-based constrained reinforcement learning using generalized control Barrier Function. *International Conference on Intelligent Robots and Systems)*. Electr Network: IEEE, 2021: 4552 – 4559.

[49] SAUNDERS W, SASTRY G, STUHLMÜLLER A, et al. Trial without error: Towards safe reinforcement learning via human intervention. *International Conference on Autonomous Agents and MultiAgent Systems*. Stockholm, Sweden: Assoc Comp Machinery, 2018: 2067 – 2069.

[50] TURCHETTA M, KOLOBOV A, SHAH S, et al. Safe reinforcement learning via curriculum induction. *ArXiv Preprint*, 2020: arXiv: 2006.12136.

[51] PENG Z H, LI Q Y, LIU C X, et al. Safe driving via expert guided policy optimization. *ArXiv Preprint*, 2021: arXiv: 2110.06831.

[52] LI Q Y, PENG Z H, ZHOU B L. Efficient learning of safe driving policy via human-ai copilot optimization. *ArXiv Preprint*, 2022: arXiv: 2202.10341.

[53] PRAKASH B, KHATWANI M, WAYTOWICH N, et al. Improving safety in reinforcement learning using model-based architectures and human intervention. *ArXiv Preprint*, 2019: arXiv: 1903.09328.

[54] SUN H, XU Z, FANG M, et al. Safe exploration by solving early terminated MDP. *ArXiv Preprint*, 2021: arXiv: 2107.04200.

[55] ALSHIEKH M, BLOEM R, EHLERS R, et al. Safe reinforcement learning via shielding. *AAAI Conference on Artificial Intelligence*. New Orleans, LA: AAAI, 2018: 2669 – 2678.

[56] HEWING L, WABERSICH K P, MENNER M, et al. Learning-based model predictive control: Toward safe learning in control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2020, 3(1): 269 – 296.

[57] JANSEN N, KONIGHOFER B, JUNGES S, et al. Safe reinforcement learning using probabilistic shields. *ArXiv Preprint*, 2018: arXiv: 1807.06096v2.

[58] LI S, BASTANI O. Robust model predictive shielding for safe reinforcement learning with stochastic dynamics. *IEEE International Conference on Robotics and Automation*. Paris, France: IEEE, 2020: 7166 – 7172.

[59] WAGENER N C, BOOTS B, CHENG C A. Safe reinforcement learning using advantage-based intervention. *ArXiv Preprint*, 2021: arXiv: 2106.09110v2.

[60] DALAL G, DVIJOTHAM K, VECERIK M, et al. Safe exploration in continuous action spaces. *ArXiv Preprint*, 2017: arXiv: 1801.08757.

[61] MO S, PEI X, WU C. Safe reinforcement learning for autonomous vehicle using monte carlo tree search. *IEEE Transactions on Intelligent Transportation Systems*, 2022, 23(7): 6766 – 6773.

[62] MAZOUCHI M, NAGESHRAO S, MODARES H. Conflict-aware safe reinforcement learning: A meta-cognitive learning framework. *IEEE/CAA Journal of Automatica Sinica*, 2022, 9(3): 466 – 481.

[63] SRINIVASAN K, EYSENBACH B, HA S, et al. Learning to be safe: Deep RL with a safety critic. *ArXiv Preprint*, 2020: arXiv: 2010.14603.

[64] THANANJEYAN B, BALAKRISHNA A, NAIR S, et al. Recovery RL: Safe reinforcement learning with learned recovery zones. *IEEE Robotics and Automation Letters*, 2021, 6(3): 4915 – 4922.

[65] BHARADHWAJ H, KUMAR A, RHINEHART N, et al. Conservative safety critics for exploration. *ArXiv Preprint*, 2020: arXiv: 2010.14497.

[66] YANG Q, SIMÃO T D, TINDEMANS S H, et al. Safety-constrained reinforcement learning with a distributional safety critic. *Machine Learning*, 2023, 112(3): 859 – 887.

[67] DUIVENVOORDEN R R, BERKENKAMP F, CARION N, et al. Constrained Bayesian optimization with particle swarms for safe adaptive controller tuning. *IFAC-PapersOnLine*, 2017, 50(1): 11800 – 11807.

[68] TURCHETTA M, BERKENKAMP F, KRAUSE A. Safe exploration in finite Markov decision processes with Gaussian processes. *ArXiv Preprint*, 2016: arXiv: 1606.04753.

[69] WACHI A, SUI Y, YUE Y, et al. Safe exploration and optimization of constrained mdps using Gaussian processes. *The 32nd AAAI Conference on Artificial Intelligence and 30th Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*. New Orleans, LA: AAAI, 2018: 6548 – 6555.

[70] KIM Y, ALLMENDINGER R, LÓPEZ-IBÁÑEZ M. Safe learning and optimization techniques: Towards a survey of the state of the art. *Lecture Notes in Computer Science*. Spain: Springer, 2021: 123 – 139.

[71] PINTO L, DAVIDSON J, SUKTHANKAR R, et al. Robust adversarial reinforcement learning. *The 34th International Conference on Machine Learning*. Sydney, Australia: PMLR, 2017, 70: 2817 – 2826.

[72] PAN X, SEITA D, GAO Y, et al. Risk averse robust adversarial reinforcement learning. *International Conference on Robotics and Automation*. Montreal, QC, Canada: IEEE, 2019: 8522 – 8528.

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

[74] VINITSKY E, DU Y, PARVATE K, et al. Robust reinforcement learning using adversarial populations. *ArXiv Preprint*, 2020: arXiv:2008.01825.

[75] LÜTJENS B, EVERETT M, HOW J P. Certified adversarial robustness for deep reinforcement learning. *ArXiv Preprint*, 2019: arXiv: 1910.12908.

[76] WU J, SIBAI H, VOROBEYCHIK Y. Certifying safety in reinforcement learning under adversarial perturbation attacks. *ArXiv Preprint*, 2022: arXiv: 2212.14115.

[77] MORIMOTO J, DOYA K. Robust reinforcement learning. *Neural Computation*, 2005, 17(2): 335 – 359.

[78] SADEGHI F, LEVINE S. CAD2RL: Real single-image flight without a single real image. *ArXiv Preprint*, 2016: arXiv: 1611.04201.

[79] LOQUERCIO A, KAUFMANN E, RANFTL R, et al. Deep drone racing: From simulation to reality with domain randomization. *IEEE Transactions on Robotics*, 2020, 36(1): 1 – 14.

[80] RAJESWARAN A, GHOTRA S, RAVINDRAN B, et al. EPOpt: Learning robust neural network policies using model ensembles. *ArXiv Preprint*, 2016: arXiv: 1610.01283.

[81] MEHTA B, DIAZ M, GOLEMO F, et al. Active domain randomization. *ArXiv Preprint*, 2019: arXiv: 1904.04762.

[82] FISAC J F, LUGOVOY N E, RUBIES-ROYO V, et al. Bridging Hamilton-Jacobi safety analysis and reinforcement learning. *International Conference on Robotics and Automation*. Montreal, QC, Canada: IEEE, 2019: 8550 – 8556.

[83] FISAC J F, AKAMETALU A K, ZEILINGER M N, et al. A general safety framework for learning-based control in uncertain robotic systems. *IEEE Transactions on Automatic Control*, 2019, 64(7): 2737 – 2752.

[84] CHOI J J, LEE D, SREENATH K, et al. Robust control barrier-value functions for safety-critical control. *The 60th IEEE Conference on Decision and Control*. Austin, TX, USA: IEEE, 2021: 6814 – 6821.

[85] HERBERT S, CHOI J J, SANJEEV S, et al. Scalable learning of safety guarantees for autonomous systems using Hamilton-Jacobi reachability. *International Conference on Robotics and Automation*. Xi'an, China: IEEE, 2021: 5914 – 5920.

[86] SELIM M, ALANWAR A, KOUSIK S, et al. Safe reinforcement learning using black-box reachability analysis. *IEEE Robotics and Automation Letters*, 2022, 7(4): 10665 – 10672.

[87] YU D, MA H, LI S, et al. Reachability constrained reinforcement learning. *The 39th International Conference on Machine Learning*. Baltimore, Maryland, USA: PMLR, 2022, 162: 25636 – 25655.

[88] YU D, ZOU W, YANG Y, et al. Safe model-based reinforcement learning with an uncertainty-aware reachability certificate. *ArXiv Preprint*, 2022: arXiv: 2210.07553.

[89] KAHN G, VILLAFLOR A, PONG V, et al. Uncertainty-aware reinforcement learning for collision avoidance. *ArXiv Preprint*, 2017: arXiv: 1702.01182.

[90] LUTJENS B, EVERETT M, HOW J P. Safe reinforcement learning with model uncertainty estimates. *ArXiv Preprint*, 2018: arXiv: 1810.08700.

[91] ZHANG J, CHEUNG B, FINN C, et al. Cautious adaptation for reinforcement learning in safety-critical settings. *ArXiv Preprint*, 2020: arXiv: 2008.06622.

[92] URPÍ N A, CURI S, KRAUSE A. Risk-averse offline reinforcement learning. *ArXiv Preprint*, 2021: arXiv: 2102.05371.

[93] ZHANG S, LIU B, WHITESON S. Mean-variance policy iteration for risk-averse reinforcement learning. *AAAI Conference on Artificial Intelligence*. Electr Network: AAAI, 2021, 35(12): 10905 – 10913.

[94] MA X, MA S, XIA L, et al. Mean-semivariance policy optimization via risk-averse reinforcement learning. *Journal of Artificial Intelligence Research*, 2022, 75: 569 – 595.

[95] SCHULMAN J, LEVINE S, ABBEEL P, et al. Trust region policy optimization. *The 32nd International Conference on Machine Learning*. Lille, France: 2015, 37: 1889 – 1897.

[96] TESAURO G, OTHERS. Temporal difference learning and td-gammon. *Communications of the ACM*, 1995, 38(3): 58 – 68.

[97] KHALIL H K. *Nonlinear Systems*. 2nd ed. New Jersey: Prentice-Hall, 1996.

[98] ROSEN J B. The gradient projection method for nonlinear programming. part i. linear constraints. *Journal of the Society for Industrial and Applied Mathematics*, 1960, 8(1): 181 – 217.

[99] BYRNE R W, RUSSON A E. Learning by imitation: A hierarchical approach. *Behavioral and Brain Sciences*, 1998, 21(5): 667 – 684.

[100] KUMAR A, ZHOU A, TUCKER G, et al. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 2020, 33: 1179 – 1191.

[101] GARNETT R. *Bayesian Optimization*. Cambridge: Cambridge University Press, 2023.

[102] SUI Y, GOTOVOS A, BURDICK J, et al. Safe exploration for optimization with Gaussian processes. *The 32nd International Conference on Machine Learning*. Lille, France: PMLR, 2015, 37: 997 – 1005.

[103] BERKENKAMP F, KRAUSE A, SCHOELLIG A P. Bayesian optimization with safety constraints: Safe and automatic parameter tuning in robotics. *Machine Learning*, 2021, 112(10): 3713 – 3747.

[104] SUI Y, ZHUANG V, BURDICK J, et al. Stagewise safe Bayesian optimization with Gaussian processes. *ArXiv Preprint*, 2018: arXiv: 1806.07555.

[105] BAUMANN D, MARCO A, TURCHETTA M, et al. GoSafe: Globally optimal safe robot learning. *International Conference on Robotics and Automation*. Xi'an, China: IEEE, 2021: 4452 – 4458.

[106] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks. *Communications of the ACM*, 2020, 63(11): 139 – 144.

[107] MITCHELL I, BAYEN A, TOMLIN C. A time-dependent Hamilton-Jacobi formulation of reachable sets for continuous dynamic games. *IEEE Transactions on Automatic Control*, 2005, 50(7): 947 – 957.

[108] CHUA K, CALANDRA R, MCALLISTER R, et al. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *ArXiv Preprint*, 2018: arXiv: 1805.12114.

[109] XIONG X, WANG J, ZHANG F, et al. Combining deep reinforcement learning and safety based control for autonomous driving. *ArXiv Preprint*, 2016: arXiv: 1612.00147.

[110] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning. *ArXiv Preprint*, 2015: arXiv: 1509.02971.

[111] ISELE D, NAKHAEI A, FUJIMURA K. Safe reinforcement learning on autonomous vehicles. *International Conference on Intelligent Robots and Systems*. Madrid, Spain: IEEE, 2018: 1 – 6.

[112] BOUTON M, KARLSSON J, NAKHAEI A, et al. Reinforcement learning with probabilistic guarantees for autonomous driving. *ArXiv Preprint*, 2019: arXiv: 1904.07189.

[113] CHEN D, JIANG L, WANG Y, et al. Autonomous driving using safe reinforcement learning by incorporating a regret-based human lane-changing decision model. *American Control Conference*. Denver, CO, USA: IEEE, 2020: 4355 – 4361.

[114] RONG J, LUAN N. Safe reinforcement learning with policy-guided planning for autonomous driving. *International Conference on Mechatronics and Automation*. Beijing, China: IEEE, 2020: 320 – 326.

[115] KRASOWSKI H, WANG X, ALTHOFF M. Safe reinforcement learning for autonomous lane changing using set-based prediction. *The 23rd International Conference on Intelligent Transportation Systems*. Electr Network: IEEE, 2020: 1 – 7.

[116] AMES A D, COOGAN S, EGERSTEDT M, et al. Control barrier functions: Theory and applications. *ArXiv Preprint*, 2019: arXiv: 1903.11199.

[117] KALWEIT G, HÜGLE M, WERLING M, et al. Interpretable multi time-scale constraints in model-free deep reinforcement learning for autonomous driving. *ArXiv Preprint*, 2020: arXiv: 2003.09398.

[118] LOPEZ P A, WIESSNER E, BEHRISCH M, et al. Microscopic traffic simulation using SUMO. *The 21st International Conference on Intelligent Transportation Systems*. Maui, HI: IEEE, 2018: 2575 – 2582.

[119] KRAJEWSKI R, BOCK J, KLOEKER L, et al. The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems. *The 21st International Conference on Intelligent Transportation Systems*. Maui, HI: IEEE, 2018: 2118 – 2125.

[120] KENDALL A, HAWKE J, JANZ D, et al. Learning to drive in a day. *ArXiv Preprint*, 2018: arXiv: 1807.00412.

[121] KINGMA D P, WELLING M. Auto-encoding variational bayes. *ArXiv Preprint*, 2013: arXiv: 1312.6114.

[122] ZHANG X, PENG Y, LUO B, et al. Model-based safe reinforcement learning with time-varying state and control constraints: An application to intelligent vehicles. *ArXiv Preprint*, 2021: arXiv: 2112.11217.

[123] MANNUCCI T, VAN KAMPEN E J, DE VISSER C C, et al. Hierarchically structured controllers for safe UAV reinforcement learning applications. *AIAA Information Systems-AIAA Infotech @ Aerospace*. Grapevine, TX, USA: AIAA, AIAA2017-0791.

[124] WANG D, FAN T, HAN T, et al. A two-stage reinforcement learning approach for multi-UAV collision avoidance under imperfect sensing. *IEEE Robotics and Automation Letters*, 2020, 5(2): 3098 – 3105.

[125] WANG F, ZHOU B, CHEN K, et al. Intervention aided reinforcement learning for safe and practical policy optimization in navigation. *ArXiv Preprint*, 2018: arXiv: 1811.06187v1.

[126] GARCIA J, FERNÁNDEZ F. Safe exploration of state and action spaces in reinforcement learning. *Journal of Artificial Intelligence Research*, 2012, 45: 515 – 564.

[127] LU X, XIAO L, NIU G, et al. Safe exploration in wireless security: A safe reinforcement learning algorithm with hierarchical structure. *IEEE Transactions on Information Forensics and Security*, 2022, 17: 732 – 743.

[128] PHAM T H, DE MAGISTRIS G, TACHIBANA R. OptLayer-Practical constrained optimization for deep reinforcement learning in the real world. *International Conference on Robotics and Automation*. Brisbane, QLD, Australia: IEEE, 2018: 6236 – 6243.

[129] GARCÍA J, SHAFIE D. Teaching a humanoid robot to walk faster through safe reinforcement learning. *Engineering Applications of Artificial Intelligence*, 2020, 88: 103360.

[130] GOUAILLIER D, HUGEL V, BLAZEVIC P, et al. Mechatronic design of NAO humanoid. *International Conference on Robotics and Automation*. Kobe, Japan: IEEE, 2009: 769 – 774.

[131] YANG T Y, ZHANG T, LUU L, et al. Safe reinforcement learning for legged locomotion. *IEEE/RSJ International Conference on Intelligent Robots and Systems*. Kobe, Japan: IEEE, 2022: 2454 – 2461.

[132] ZHU X, KANG S, CHEN J. A contact-safe reinforcement learning framework for contact-rich robot manipulation. *IEEE/RSJ International Conference on Intelligent Robots and Systems*. Kyoto, Japan: IEEE, 2022: 2476 – 2482.

[133] RAY A, ACHIAM J, AMODEI D. Benchmarking safe exploration in deep reinforcement learning. *Open AI*, 2019: https://openai.com/research/safety-gym.

[134] YUAN Z, HALL A W, ZHOU S, et al. Safe-control-gym: A unified benchmark suite for safe learning-based control and reinforcement learning in robotics. *IEEE Robotics and Automation Letters*, 2022, 7(4): 11142 – 11149.

[135] ZHANG L, ZHANG Q, SHEN L, et al. Saferl-kit: Evaluating efficient reinforcement learning methods for safe autonomous driving. *ArXiv Preprint*, 2022: arXiv: 2206.08528.

[136] FU J, KUMAR A, NACHUM O, et al. D4rl: Datasets for deep data-driven reinforcement learning. *ArXiv Preprint*, 2020: arXiv: 2004.07219.

[137] QIN R J, ZHANG X, GAO S, et al. Neorl: A near real-world benchmark for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 2022, 35: 24753 – 24765.

作者简介:

张昌昕　博士研究生, 目前研究方向为安全强化学习、机器人和智能驾驶, E-mail: changxzhang@163.com;

张兴龙　副教授, 博士, 在Automatica, IEEE Transactions汇刊等国际期刊和会议发表论文40余篇, 目前研究方向为滚动时域强化学习及其在无人系统中的应用, E-mail: zhangxinglong18@nudt.edu.cn;

徐　昕　教授, 博士, 在国际期刊和会议发表学术论文200余篇, 出版书籍4本, 目前研究方向包括智能控制、强化学习、近似动态规划、机器学习、机器人和智能驾驶, E-mail: xuxin_mail@263.net;

陆　阳　博士研究生, 目前研究方向为强化学习及其在无人系统中的应用, E-mail: luyang18@nudt.edu.cn.