基于鲁棒交叉熵与梯度优化的安全强化学习方法

周娴玮,张 锟,叶 鑫

(华南师范大学 软件学院,广东 佛山 538200)

摘 要: 智能体在复杂环境下执行任务时,如何保证安全性和效率性是一个很大的难题。传统强化学习方法解决智能体决策问题时采用无模型的强化学习,利用大量数据不断试错寻找最优策略,忽略了智能体的训练成本和安全风险,因此无法有效保证决策的安全性。为此,在模型预测控制框架下对智能体动作添加安全约束条件,设计安全强化学习算法获得最安全的动作控制序列。同时,针对交叉熵方法存在计算量大与效率低、梯度优化方法存在着陷入局部最优的问题,结合鲁棒交叉熵与梯度优化方法优化动作控制序列,以提升算法安全性和求解效率。实验表明,所提方法相较于鲁棒交叉熵法能有效提升收敛速度,相较于其他优化算法在不损失较多性能的前提下安全性能最优。

关键词: 强化学习;鲁棒交叉熵;梯度优化;安全性

DOI: 10. 11907/rjdk. 231853 **开放科学(资源服务)标识码(OSID):**

中图分类号: TP391.4 **文献标识码:** A **文章编号:** 1672-7800(2024)009-0143-07

Safe Reinforcement Learning Method Based on Robust Cross-Entropy and Gradient Optimization

ZHOU Xianwei, ZHANG Kun, YE Xin

(*School of Software*, *South China Normal University*, *Foshan* 538200, *China*)

Abstract: Ensuring security and efficiency when intelligent agents perform tasks in complex environments is a major challenge. Traditional reinforcement learning methods use model free reinforcement learning to solve intelligent decision-making problems, constantly trial and error to find the optimal strategy using a large amount of data, ignoring the training cost and security risks of the agent, and therefore cannot effectively ensure the safety of decision-making. To this end, safety constraints are added to the actions of intelligent agents in the model predictive control framework, and a safety reinforcement learning algorithm is designed to obtain the safest action control sequence. At the same time, in response to the problems of high computational complexity and low efficiency in the cross entropy method, as well as the problem of falling into local optima in the gradient optimization method, a combination of robust cross entropy and gradient optimization methods is used to optimize the action control sequence to improve algorithm safety and solving efficiency. The experiment shows that the proposed method can effectively improve the convergence speed compared to the robust cross entropy method, and has the best safety performance compared to other optimization algorithms without sacrificing much performance.

Key Words: reinforcement learning; robust cross-entropy; gradient optimization; safety

0 引言

经典的强化学习智能体导航方法重点在如何高效探索环境,通过智能体不断试错寻找智能体控制的最优策略,因此在该模式下只要能提高控制性能就允许智能体执行任何动作[1]。智能体探索出各种动作后根据环境反馈的奖励优化选择动作,最终得出一类奖励最大的策略(状态—动作序列)[2]。然而,这一过程包含大量试错与危险动作,在实际环境中应用该算法可能导致装置损坏甚至人

收稿日期: 2023-08-10

扫描二维码阅读全文:

基金项目: 广东省基础与应用基础研究基金项目(2020A1515110783);广东省企业科技特派员项目(GDKTP2020014000);佛山市高层次派驻人才项目(303475)

作者简介: 周娴玮(1980-),男,博士,华南师范大学软件学院讲师、硕士生导师,研究方向为机器人自动化技术;张锟(1998-),男,CCF学生会员,华南师范大学软件学院硕士研究生,研究方向为安全强化学习;叶鑫(1996-),男,CCF学生会员,华南师范大学软件学院硕士研究生,研究方向为强化学习。本文通讯作者:张锟。

员受伤。因此,强化学习需要限制这些对高安全领域(工业机器人、自动驾驶等)而言较危险的动作,于是要求研究人员不仅要关注长期奖励最大化,还要关注训练中智能体行为,避免其带来损伤。安全强化学习的出现使智能体在高安全性领域能控制自己的探索行为,即在满足安全约束的情况下通过最大化期望回报值得到最优策略[3]。安全强化学习能保证智能体在训练期间对周围环境和智能体本身的安全性,但目前方法存在计算量大、效率低和容易陷入局部最优的问题,导致算法性能不佳、安全性较低。

为此,本文在模型预测控制框架下对智能体动作添加安全约束条件,设计安全强化学习算法,从而获得最安全的动作控制序列。同时,结合鲁棒交叉熵与梯度优化方式优化动作控制序列,以解决交叉熵方法存在计算量大与效率低、梯度优化方法存在着陷入局部最优的问题,从而提升算法的安全性和求解效率,致力于提升智能体在训练过程中的效率和安全性[4]。本文主要贡献如下:①研究了在模型预测控制(Model predictive control, MPC)和基于模型的安全强化学习背景下的智能体动作规划与优化问题;②提出一种结合梯度优化和鲁棒性交叉熵的安全强化学习算法;③在 Safety Gym 平台上,将本文方法与交叉熵(Cross-Entropy Method, CEM)方法、随机打靶法(Random Shooting)、鲁棒交叉熵法(Robust Cross-Entropy, RCE)等安全强化学习算法,在任务性能、安全性能等方面进行比较。

1 相关工作

1.1 安全强化学习

目前,强化学习领域已积累了众多研究基础。Schulman 等[5]提出的 TRPO(Trust Region Policy Optimization)算法限制了更新步长,构建了一个用于约束强化学习的信任区域,在数学上证明了该算法会收敛到局部最优或全局最优。Achiam 等[6]提出 CPO(Constrained Policy Optimization)方法继承了 TRPO 的思想,用于解决安全约束条件下的智能体运动规划问题,该方法通过奖励函数和安全约束条件塑造智能体动作行为,保证最终收敛到满足安全约束条件的最优策略,但只能近似满足约束条件,如图 1 所示。Tessler 等[7]提出奖励值受限的 RCPO(Reward Constrained Policy Optimization)设置约束避免智能体找到奖励漏洞,解决了 CPO 方法的部分误差与问题。Yu 等[8]构造 Lyapunov Function,将控制理论引入强化学习,保证策略在训练和部署中的全局安全性。

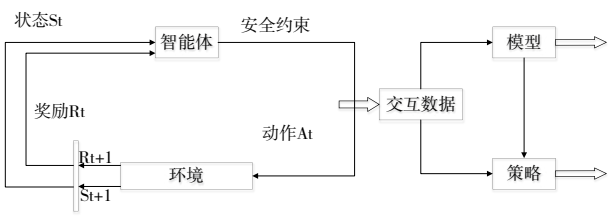


Fig. 1 Safe reinforcement learning model

图 1 安全强化学习模型

此外,还有学者通过修改探索过程解决安全性问题,但非常依赖于安全层的预测;或提出 CAP 算法,利用保守性的成本约束条件使其适应真实环境,以保证在训练阶段智能体的安全性;或基于受约束交叉熵优化方法提出一种无模型安全强化学习算法,以解决有限时域内的受约束规划问题,并对算法的收敛性进行严格证明;或针对安全约束未知的 CMDP 提出一种策略优化算法,通过扩张安全区域学习安全约束后在安全区域内最大化累积奖赏。

虽然上述安全强化学习方法都能近似解决训练时的安全问题,为提升智能体安全性提供思路,但仍存在训练效率过慢、收敛效果不佳、未能严格解决智能体探索过程中的安全等问题[8]。在强化学习的基础上,安全强化学习的目标是找到一个策略,在满足预先设定的一系列安全约束基础之上,最大化智能体在无穷时域内的累积奖赏的期望值。在受限的马尔可夫过程框架下,一种常用的安全约束要求智能体在无穷时域内的累积代价后小于某个阈值[9]。具体数学表达式为:

$$E\sum_{t=0}^{\infty}\gamma^t C_i(s_t) \leqslant d_i',\ \forall i \in \{1,\ 2..,\ m\} \tag{1}$$

式中:C_i 表示安全约束值;s_t 表示智能体状态值;d_i' 表示智能体动作约束阈值。

然而,此类约束物理意义不够明晰,无法严格保证智能体不到危险区域。为此,本文要求智能体的累积代价函数必须小于某个阈值,即此时安全强化学习的数学模型为:

$$\pi^* = argmaxE\sum_{t=0}^{\infty}\gamma^t r\left(s_t, \pi\left(s_t\right)\right) \tag{2}$$

$$s.t.\ C_i\left(s_t\right) \leqslant di$$

式中:$\forall i \in \{1,\ 2..,\ m\},\ \forall t \geqslant 0$;$C_i(s)$ 为受限马尔可夫决策过程的约束函数;di 为每一个约束函数对应的阈值;r 为奖励函数;s_t 代表状态值;π^* 为待求解的最佳策略动作序列;γ^t 为折扣因子。

经过与环境交互学到的模型不仅能预测状态轨迹与该状态轨迹的累积奖赏,还能预测该状态轨迹的安全性[10]。在应用强化学习训练策略时,相应的奖励函数和代价函数通常是人为设计。以移动智能体控制规划下的导航任务为例,人们通常根据智能体与目标点的距离设计奖励函数,根据智能体与危险区域的距离设计代价函数。因此,此类已知奖励函数和代价函数的假设在实际应用中较为合理且容易满足[11]。

1.2 基于鲁棒交叉熵的模型预测与控制方法

在基于模型的强化学习和模型预测控制中,从真实环境采集的数据学习模型近似系统动力学模型,然后利用模型进行控制规划得到最优控制序列[12]。模型预测控制属于优化和控制领域的交叉领域,目的是找到最优控制,具体原理为在最优控制策略基础上叠加预测控制,也叫滚动

优化过程,即预测未来几步的系统状态,根据未来预测状态求解最优控制量,并最终选择最近一步的控制量,作用在下一个周期后再重复计算。在实际中,MPC每步都会重复搜索整个预测过程以检测环境变化,这一过程通常基于鲁棒交叉熵(Robust Cross-Entropy, RCE)方法[13]。首先,从初始样本的高斯分布中随机采样N个样本,计算每个样本的目标函数值(奖励函数值和代价函数值),选出目标函数值最高的$N_b(N_b < N)$个样本,并评估这些样本的安全性能和任务性能,以决定可行决策集合,由此得到任务性能和安全性能都较为优秀的控制序列。其次,参与模型进一步的迭代训练。此外,每次迭代时根据目标函数值选取一定数量的候选序列更新高斯分布参数,以获得预期累计奖励最高的行为决策动作[14,15]。

然而,这种方式实际中存在弊端,在交叉熵排序评估后选择直接舍弃后面的次优动作,且重复搜索解决优化问题时计算量巨大[16]。实际上,可利用梯度优化和交叉熵相结合的方法来指导最优化控制过程中的搜索过程,而非直接对无序的动作序列进行简单抽样和排序,因此本文提出基于鲁棒交叉熵和梯度优化的安全强化学习算法。

1.3　梯度优化

梯度优化的智能体控制规划方法通常利用奖励函数的反向传播导数,采用梯度下降方式迭代更新动作序列[17]。本文采用随机梯度下降,具体迭代过程为:

$$\bar{a}_{0:T}^{k} \coloneqq \bar{a}_{0:T}^{k-1} + \alpha \nabla_{\bar{a}} \bar{\mathrm{R}}\left(\bar{a}_{0:T}^{k-1}, \bar{s}_{0:T}^{k}\right) \tag{3}$$

式中:T表示时间范围;a表示动作;s表示状态;α为学习率。

梯度下降方法迭代更新动作序列的最大缺点是在优化过程中容易陷入局部最优,且迭代次数较多,在解空间中的搜索过程十分繁杂。因此,在基于强化学习的模型预测控制中可能收敛到非最优策略,还可能存在梯度优化和梯度丢失的问题[18,19]。强化学习要保证被控制物体的安全,就需要改变传统建模方法,构建出融合智能体控制与强化学习的安全模型[20]。为此,本文提出了基于鲁棒交叉熵和梯度优化的安全强化学习方法。

2　本文算法

考虑在智能体模型预测控制方法中将鲁棒交叉熵法与随机梯度优化相结合以指导智能体运动规划。该方法相较于传统交叉熵方法,采用了更安全的鲁棒交叉熵法,并考虑了在智能体运动规划阶段的计算量大的问题,同时使用梯度下降思想指导安全动作优化过程,产生更准确的智能体动作控制序列,以加速交叉熵优化过程[21,22]。

2.1　基于安全模型的模型预测与控制算法

本文使用模型预测控制(Model Predictive Control, MPC)作为安全强化学习的基本模型框架,基本原理为利用模型预测系统在未来某一个时间段的动作表现$\chi = (a_0, a_1, \cdots a_T)$,并将第一个动作引入模型,通过系统检测获得反馈和新的观测值,然后执行类似的优化过程。此外,利用MPC在控制优化方面的特点引入安全约束条件,将原本简单的规划问题转化为带安全约束的优化问题[23]。

基于MPC框架的强化学习优势是样本效率较高、可提供虚拟样本。通过智能体与环境交互得到的真实样本训练模型,然后利用模型产生大量虚拟样本辅助训练,从而大幅提升样本效率。具体为,首先将采集的数据学习模型来近似系统动力学模型$f\left(s_t, a_t\right)$;然后利用模型控制规划得到动作序列方法,主要思想是通过环境交互构建动力学模型和代价模型,并利用模型控制规划[24]。其中,动力学模型根据当前时刻状态和当前动作预测下一时刻状态,即$(s_t, a_t) \rightarrow s_{t+1}$,如图2所示。

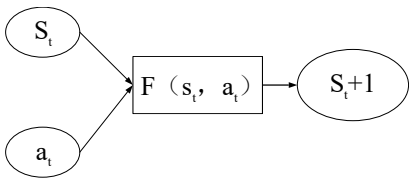


Fig. 2　Dynamic model

图2　动力学模型

本文代价模型$c\left(s_{t+1}\right)$采用一种二元分类模型的决策树来近似代替,模型动作的安全的分类标签为:

$$c\left(s_{t+1}\right) = \begin{cases} 0 & safe \\ 1 & unsafe \end{cases} \tag{4}$$

首先经过分类器将整个数据分离到两个缓冲区,一个用于安全数据,另一个用于不安全数据,以控制训练安全数据与不安全数据的占比。然后通过学习到的动力学模型和代价模型代替实际系统动态,求解出在M步内满足安全约束的最优控制序列。本文目标是利用模型求解满足安全约束条件下的智能体运动控制最优策略,此时受约束的优化问题的数学表达式为:

$$\chi = argmaxE \sum_{t=0}^{\infty} \gamma^t r\left(s_{t+1}\right) \tag{5}$$

$$s.t. \quad s_{t+1} = f\left(s_t, a_t\right),\ c\left(s_{t+1}\right) = 0 \tag{6}$$

式中:$\forall t \in \{0, 1, \cdots, T-1\}$;$\gamma$为折现因子;$r\left(s_{t+1}\right)$为奖励函数;$\chi$为动作序列;$c\left(s_{t+1}\right) = 0$为代价函数;$s_{t+1}$为下一时刻状态值;$s_t$为当前时刻状态值;$f$为动力学模型;$c$为代价模型。

2.2　基于鲁棒交叉熵与梯度优化的强化学习算法

本文使用鲁棒交叉熵法和梯度优化来解决安全约束优化问题,但由于交叉熵方法计算量巨大且容易造成资源浪费,梯度下降容易陷入局部最优解。为此,在鲁棒交叉熵方法的步骤中穿插梯度优化过程,利用梯度下降优点改进动作序列优化过程,从而产生更精确的智能体控制动作序列,以提升算法实际效率,实现更快的更新迭代控制动作。

假设$f\left(s_t, a_t\right)$表示学习到的动力学模型,r表示学习后

的奖励模型，a_t 表示当前 t 时刻下的动作，T 表示计划视界，$\mathcal{N}(\mu_{0:T}^b, \Sigma_{0:T}^b)$ 表示交叉熵法中的抽样分布结果，初始化后为 $\mathcal{N}(\mu_{0:T}^t, \Sigma_{0:T}^t = I)$。

每次迭代开始时，利用一种轨迹采样器方法（G）对动作序列进行随机采样：

$$\{(a_0^b, a_1^b, \cdots a_T^b)\}_{g=1}^G \sim \mathcal{N}(\mu_{0:T}^b, \Sigma_{0:T}^b) \tag{7}$$

式中：a_0^b 表示第 b 次采样迭代第一个时刻智能体的动作；g 表示采样的序列。

利用学习到的当前时刻动力学模型 $f(s_t, a_t)$ 和当前奖励函数模型 $r(\chi;\ s_t^b)$ 评估得到当前动作序列的预计累积奖励值 R_g^b、累计代价值 c_g^b。为了提升算法鲁棒性，使算法安全性能更优，首先依据累积的代价值排序，选择安全性能更好的动作序列，再综合奖励值选择序列进入下一步骤，具体奖励值和代价值评估标准为：

$$R_g^b = \sum_{t=1}^{T} \gamma^b r(\chi;\ s_t^b) \qquad s_t^b = f(s_{t-1}^b, a_{t-1}^b) \tag{8}$$

$$c_g^b = \sum_{t=1}^{T} \beta^b maxc(\chi, s_t^b) \quad \forall_g = 1, \cdots, G \tag{9}$$

式中：γ^b、β^b 均为折扣因子；b 为对 RCE 迭代的索引；R_g^b 表示累计奖励值；c_g^b 表示累计代价值；f 为动力学模型；s_t^b 为当前状态值。

本文实验梯度下降步数 $M = 1$，迭代采样后为了提升算法实际效率，选择保留前一次迭代过程中满足安全性评估条件下的前 K 个规划动作序列，并将采样过程视为梯度优化过中的初始情况，对以上采样得到的动作序列作 M 步的随机梯度优化。具体数学表达式为：

$$(a_0^b, a_1^b, \cdots a_T^b)_g^{m+1} \leftarrow (a_0^b, a_1^b, \cdots a_T^b)_g^m - \alpha \nabla_{a_{0:T}^b} R_g^b \tag{10}$$

$$\forall_g = 1, \cdots, G \quad j = 1, \cdots, M$$

然后，依据梯度优化动作更新分布的 $\mathcal{N}(\mu_{0:T}^{b+1}, \Sigma_{0:T}^{b+1})$ 参数，匹配出前 k 个更新后的动作序列，*Safe* 序列即为筛选安全性能最优的动作序列。

$$\mu_{0:T}^{b+1} \leftarrow Safe(\{(a_0^b, a_1^b, \cdots a_T^b)_k\}_{k=1}^K\} \tag{11}$$

$$\Sigma_{0:T}^{b+1} \leftarrow Var(\{(a_0^b, a_1^b, \cdots a_T^b)_k\}_{k=1}^K\} \tag{12}$$

经上述过程，在 B 次迭代后算法返回具有最佳安全性能前提下的最高累计奖励动作序列。具体算法步骤如下：

算法 1 RCE+Grad

输入：初始采样分布参数 $\mathcal{N}$，样本数目 m，精英样本数目 K，初始状态 $\boldsymbol{s}_0$。

输出：奖励最高的动作 χ。

1. 初始化：初始采集的数据 d。

2. 训练动力学模型 $f(s_t, a_t)$，奖励模型 $r(\chi_i;\ s_t^b)$，代价模型 $c(\chi_i, s_t^b)$。

3. for m 轮训练 do。

4. 从初始环境样本中抽取 n 个动作序列样本。

5. for *rce* 迭代 t=1 to t do。

6. for 动作序列随机梯度下降步长 j=1 do。

7. $s_t^b = f(s_{t-1}^b, a_{t-1}^b)$。

8. 根据式（8）、式（9）计算奖励值 R_g^b 和代价值 c_g^b。

9. 对动作序列 χ_i=$(a_0^b, a_1^b, \cdots a_T^b)\}_{g=1}^G$ 执行一步的随机梯度下降。

10. end。

11. 根据代价值选择出可行解集 $\psi = \{\chi_i\}$。

12. if 可行解集 ψ 是空集。

13. 对 χ_i 进行代价值升序排列，选取前 k 个的样本得到 safe 序列。

14. else。

15. 对 ψ 中元素进行奖励值降序排序，选取前 k 个样本。

16. 根据前 k 个动作序列更新 $\mathcal{N}(\mu_{0:T}^b, \Sigma_{0:T}^b)$ 分布。

17. end。

18. 获得最大奖励的动作序列。

19. 执行最大奖励动作序列中的第一个动作。

20. end。

首先对动作控制序列执行一步随机梯度下降，根据式（9）估计代价，选择满足安全约束的可行解集合；然后对可行解集中的解进行排序，选择前 k 个样本计算下一次迭代的抽样分布参数，如果所有样本都至少违反过一次约束，从样本集中抽取前 K 个代价值最低的样本。

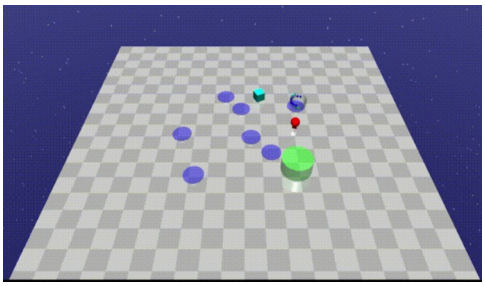
本文所提算法在计算奖励期望的同时考虑了最坏的代价，以最大代价最小为目标寻找可行解集，并直接对动作序列进行优化，主要用于解决智能体在训练过程中遇到的安全问题。因此，基于鲁棒交叉熵和梯度优化的安全强化学习方法可在受限的环境下探索规划智能体的路径，避免智能体在训练过程中遭受危险，以保证获取到一个最大的长期奖励回报。

3 实验结果与分析

3.1 实验环境

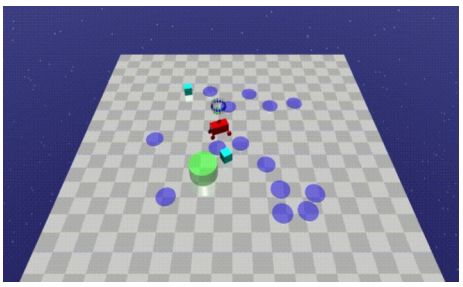
本文实验在 OpenAI 开发的 Safety Gym 环境中评估所提算法的安全性与任务性能。实验的训练和测试过程中都使用了两种场景，如图 3 所示（彩图扫 OSID 可见，下同）。每个实验场景包含一种机器人，必须在避免碰撞到障碍物的前提下在集群环境中到达目标圈（图 3 绿色柱状物）。图 3（a）中红色圆点为 safety gym，包含一个简单模拟机器人；蓝色圆圈为障碍物，模拟现实危险区域；绿色柱状物体为目标区域。图 3（b）中红色物体为定制的一种四轮汽车模型，前、后轮装有驱动装置以保证汽车模型在各个方向可以行行驶，两种机器人都有二维连续动作空间和观察空间。

本文选择 Safety Gym 环境中的导航任务，当机器人按照要求移动到指定的目标区域后（绿色柱状体），目标位置随机重置一次，环境中其余障碍设置则在一轮训练中保持不变，一轮训练结束后整个布局会被重置一次。机器人在环境中移向目标点会获得少量的奖励，到达目标点后会获



（a）　Point goal

（a）　点目标



（b）　Car goal

（b）　汽车目标

Fig. 3　Simulation environment

图3　仿真环境

得 $r_t = 1$ 的任务奖励。实验中使用了仿真平台的 Hazards、Vases 两种安全约束元素，分别代表图 3 蓝色圆圈区域和青色立方体。Hazards 为训练过程中需要避开的危险区域，Vases 为初始静止但触碰后可移动的障碍物花瓶。机器人进入危险区域或接触到花瓶，即在当前时刻智能体违反安全约束，此时代价函数 $c_t = 1$。

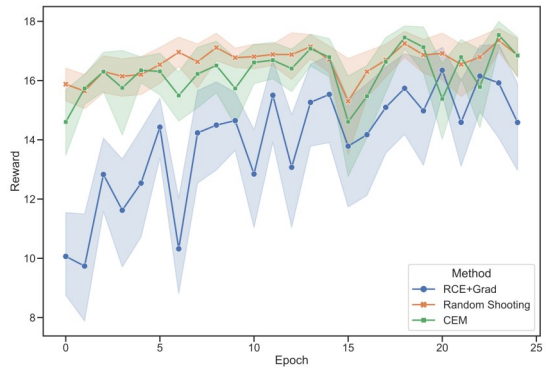
3.2　实验设置与评价指标

实验中采用 MPC 作为模型控制框架，将交叉熵方法（Cross-Entropy，CEM）、随机射击（Random Shooting，Random）、RCE+Grad 作为比较方法，以分析所提算法的任务性能和安全性能。此外，设置 RCE、RCE+Grad 方法在训练过程中模型的交互时间、奖励变化、实际收敛速率，根据实验效果分析 RCE+Grad 方法的优缺点和适用范围。

各安全强化学习方法（CEM、Random shooting、RCE+Grad）在 MPC 中样本数目设置为 500，对其进行初始样本采样。训练期间采用随机梯度下降方式，在点目标实验中进行 b=25 次迭代，在汽车目标实验中进行 b=40 次迭代。在比较 RCE 与 RCE+Grad 方法时，增加了点目标实验环境中的障碍物数目，在相同交互次数的情况下总奖励变化情况为标准测试两种算法的实际收敛效果。实验从每轮训练过程中的累计奖励值和累计代价值两个角度比较不同方法的优劣。其中，累计奖励描述算法任务性能，累计代价值表示每轮中违反安全约束的次数，最后比较训练期间的总成本。此外，还需比较收敛所需的总训练时间，以反映算法实际收敛速率。

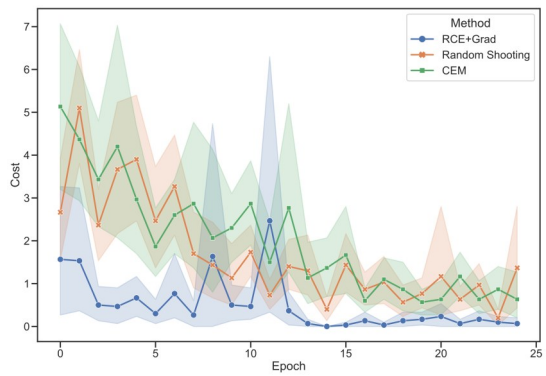
3.3　结果分析

由于安全强化学习中的智能体是通过试错的方法进行学习，并且在本文算法中假设没有先验知识，只通过环境交互学习动力学模型和代价模型。因此，在训练早期不可避免的会产生一些违反安全约束的数据。本文算法与 CEM 和 Random shooting 的比较结果如图 4 所示。其中，横坐标代表训练过程中模型与环境交互次数，纵坐标 Reward 表示奖励值，衡量算法任务性能情况；纵坐标 Cost 代表代价值，衡量算法安全性能。



（a）　Point goal reward value

（a）　点目标奖励值



（b）　Point goal cost value

（b）　点目标代价值

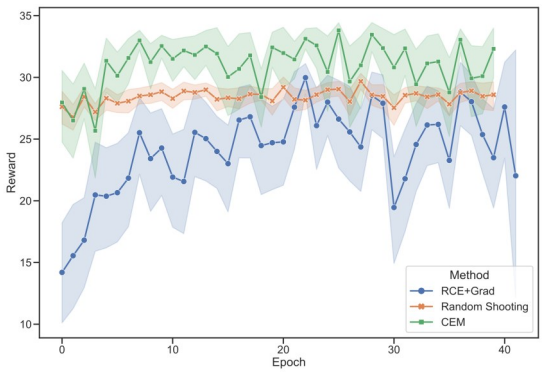
Fig. 4　Point goal experiment

图4　点目标实验

在目标实验中，对具有相同超参数但不同随机种子的 3 个实验取奖励值和代价值的平均值。实验表明，CEM、Random shooting 算法在收敛后奖励值相较于 RCE+Grad 方法几乎相同，本文方法在性能上损失并不多。在安全性方面，训练开始时本文算法快速学习了底层约束函数，随模型训练时间增长代价值不断降低，整体算法代价值远低于其他算法，整个训练阶段代价值皆低于 1，有效避免了训练过程中探索不安全的行为。实验证明，尽管本文方法在收敛速率和任务性能方法略低于其他两种算法，但安全性能更优秀。

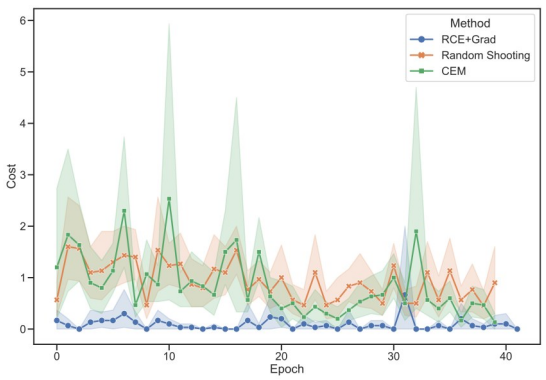
图 5 展示的是汽车目标实验。由于汽车实验目标较大，在面对障碍物较多的环境过程中学习难度较大，因此

初始训练任务性能增长较慢。在加入梯度优化方法后，算法在收敛速率效果上具有一定提升，尤其在汽车实验中的提升相较于目标实验更明显，但训练到 20 000 步左右时，算法在汽车实验任务中的性能不够稳定，可能是受到梯度优化算法的影响，但在 28 000 步左右达到收敛，最终奖励值与其他两种算法相差不大。在代价值方面，本文方法的代价值曲线远远低于其他算法，证明该算法在汽车实验中同样具有良好的安全性能。



（a） Car goal reward value

（a） 汽车目标奖励值



（b） Car goal cost value

（b） 汽车目标代价值

Fig. 5 Automotive goal experiment

图 5 汽车目标实验

通过实验发现，算法在训练初期代价值较大，经过模型训练智能体一段时间后代价值逐步降低。为此，本文通过一些历史不安全数据或模拟器预训练智能体。训练初始阶段，在保证对智能体的损害较小的前提下，刚训练就可获得较为理想的安全性能，最终将训练安全决策较佳的智能体部署到实际应用中。表 1 展示的是训练过程中安全约束条件违反情况，通过累计违反安全约束条件的次数反映算法安全性，表中数字代表训练期间步违反安全约束的总数，点目标实验总训练步数为 25 000 次，汽车目标实验总步数为 40 000 次。由此可知，本文算法在训练期间代价远低于其他算法，证明了 RCE+Grad 算法能以更安全的方式探索环境，实现了安全强化学习安全性的目标。

图 6 展示的是加入梯度优化前后两种算法的奖励值。

Table 1 Number of violations of safety constraints during training

表 1 训练中违反安全约束次数

方法任务	RCE+Grad	CEM	Random shooting
点目标实验	**98.0**	531.0	440.0
汽车目标实验	**48.0**	485.0	389.0

其中，横坐标表示强化学习与环境交互的回合数；纵坐标表示算法奖励值，以衡量算法任务性能；实线代表经过 3 次重复比较试验后的奖励值平均值；浅色区域代表一个标准差方位内的面积。此外，本文实验环境相较于点目标实验加入了更多障碍物，以进一步测试算法收敛性能。

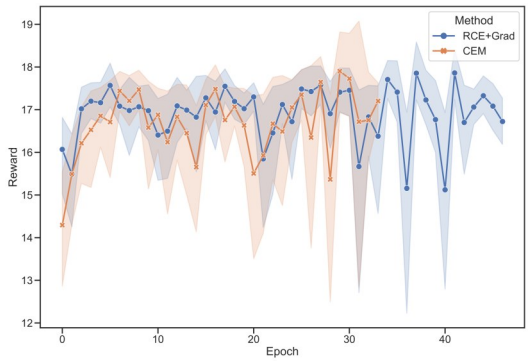


Fig. 6 Comparison between RCE+Grad method and RCE method

图 6 RCE+Grad 方法与 RCE 方法比较

由此可知，在同样环境下 RCE+Grad 算法在经过 80 轮训练后达到收敛，而 RCE 算法需要经过 120 轮训练，多次实验证明了 RCE+Grad 算法在训练前期收敛速度明显优于未加入梯度优化的 RCE 算法。同时，经过 350 轮训练后 RCE+Grad 算法计算得到的长期累计奖励值相较于普通 RCE 算法有所提高，原因为梯度优化能加速算法选择高性能的动作序列过程，同时保留了安全性序列。在未使用梯度优化方法指导安全强化学习算法中，动作序列采样过程需要多次重复迭代、排序，导致计算量过大。RCE+Grad 方法将 CEM 步骤和梯度下降穿插在样本上，以局部改进每个动作的控制规划。

综上，模型结合梯度能产生更精细的安全动作序列，可快速更新交叉熵采用分布，而非重新采样所有控制规划序列，既保证了良好的任务性能，又保留了算法在训练期间的安全约束，使其能在安全性能上表现更佳。

4 结语

本文研究了模型预测控制和基于模型强化学习背景下的安全强化学习算法，该算法无需系统动力学模型和约束模型作为预先假设，直接从环境交互中学习模型。首先，利用梯度优化思想改进鲁棒交叉熵方法，提升了传统交叉熵算法的收敛速度。然后，在 Safety Gym 环境下将本文算法与 CEM 和 Random shooting 进行比较，证实了本文方法在任务性能损失不多的情况下，安全性方面远优于交

叉熵方法和随机打靶法。

未来,将在考虑长远的安全约束的情况下,进一步改进代价模型进行以获得更优的安全性能。

参考文献:

[1] PAREKH D, PODDAR N, RAJPURKAR A, et al. A review on autonomous vehicles: progress, methods and challenges[J]. Electronics, 2022, 11(14): 2162.

[2] MOERLAND T M, BROEKENS J, PLAAT A, et al. Model-based reinforcement learning: a survey [DB/OL]. https://arxiv. org/abs/2006. 16712.

[3] YANG Y, JIANG Y, LIU Y, et al. Model-free safe reinforcement learning through neural barrier certificate[J]. IEEE Robotics and Automation Letters, 2023, 8(3): 1295-1302.

[4] BRUNKE L, GREEFF M, HALL A W, et al. Safe learning in robotics: from learning-based control to safe reinforcement learning[J]. Annual Review of Control, Robotics, and Autonomous Systems, 2022, 5: 411-444.

[5] SCHULMAN J, LEVINE S, MORITZ P, et al. Trust region policy optimization[DB/OL]. https://arxiv. org/abs/1502. 05477.

[6] ACHIAM J, HELD D, TAMAR A, et al. Constrained policy optimization [C]// International Conference on Machine Learning, 2017: 22-31.

[6] TESSLER C, MANKOWITZ D J, MANNOR S. Reward constrained policy optimization[DB/OL]. https://arxiv. org/abs/1805. 11074.

[7] CHOW Y, NACHUM O, DUENEZ-GUZMAN E, et al. A lyapunov-based approach to safe reinforcement learning[DB/OL]. https://arxiv. org/abs/1805. 07708.

[8] YU H, XU W, ZHANG H. Towards safe reinforcement learning with a safety editor policy[J]. Advances in Neural Information Processing Systems, 2022, 35: 2608-2621.

[9] BRUNKE L, GREEFF M, HALL A W, et al. Safe learning in robotics: from learning-based control to safe reinforcement learning[J]. Annual Review of Control, Robotics, and Autonomous Systems, 2022, 5: 411-444.

[10] KAMRAN D, SIMÃOR T D, YANG Q, et al. A modern perspective on safe automated driving for different traffic dynamics using constrained reinforcement learning[C]// 2022 IEEE 25th International Conference on Intelligent Transportation Systems, 2022: 4017-4023.

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

[12] KAMRAN D, SIMÃOR T D, YANG Q, et al. A modern perspective on safe automated driving for different traffic dynamics using constrained reinforcement learning[C]// 2022 IEEE 25th International Conference on Intelligent Transportation Systems, 2022: 4017-4023.

[13] WEN M, TOPCU U. Constrained cross-entropy method for safe reinforcement learning[J]. IEEE Transactions on Automatic Control, 2021, 66(7):3123-3137.

[14] MOOS J, HANSEL K, ABDULSAMAD H, et al. Robust reinforcement learning: a review of foundations and recent advances [J]. Machine Learning and Knowledge Extraction, 2022, 4(1): 276-315.

[15] WANG Y, MA X, CHEN Z, et al. Symmetric cross entropy for robust learning with noisy labels[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision,2019: 322-330.

[16] ARROYO J, MANNA C, SPIESSENS F, et al. Reinforced model predictive control (RL-MPC) for building energy management[J]. Applied Energy, 2022, 309: 118346.

[17] TOBAJAS J, GARCIA-TORRES F, RONCERO-SÁNCHEZ P, et al. Resilience-oriented schedule of microgrids with hybrid energy storage system using model predictive control[J]. Applied Energy, 2022, 306: 118092.

[18] ZHU F, GE Y Y, LING X H, et al. Model-free safe reinforcement learning method based on constrained Markov decision processes[J]. Journal of Software, 2022,33(8):3086-3102.
朱斐,葛洋洋,凌兴宏,等. 基于受限MDP的无模型安全强化学习方法[J]. 软件学报,2022,33(8):3086-3102.

[19] DAI S S, LIU Q. Action constrained deep reinforcement learning based safe automatic driving method[J]. 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[J]. 计算机科学,2021,48(9):235-243.

[20] ZHANG H R, ZHAO C H, DING J L. Robust safe reinforcement learning control of unknown continuous-time nonlinear systems with state constraints and disturbances[J]. Journal of Process Control, 2023, 128: 103028.

[21] ZHOU P W, XU Z H, ZHU X P, et al. Safe reinforcement learning method integrating process knowledge for real-time scheduling of gas supply network[J]. Information Sciences,2023,633: 280-304.

[22] ZHAO Q Y, ZHANG Y, LI X D. Safe reinforcement learning for dynamical systems using barrier certificates[J]. Connection Science, 2022, 34(1):2822-2844.

[23] LIU S H, LIU L J, YU Z. Safe reinforcement learning for affine nonlinear systems with state constraints and input saturation using control barrier functions[J]. Neurocomputing,2023,518: 562-576.

[24] RAFAEL B, BALÁZS K, IVAN S, et al. Dynamic stochastic electric vehicle routing with safe reinforcement learning [J]. Transportation Research Part E,2022,157: 102496.

(责任编辑:刘嘉文)