面向节能型驾驶辅助系统的安全强化学习

Habtamu Hailemichael * **Beshah Ayalew** * **Lindsey Kerbel** *
Andrej Ivanco ** **Keith Loiselle** **

*汽车工程，克莱姆森大学，美国南卡罗来纳州格林维尔 *29607*（*hhailem, beshah, lsutto2*）*@clemson.edu*。
**艾里逊变速器公司，美国印第安纳州印第安纳波利斯 *One Allison Way*，*IN 46222*（*andrej.ivanco, keith.loiselle*）*@allisontransmission.com*

摘要： 基于强化学习（RL）的驾驶辅助系统旨在利用来自实际应用的经验数据，不断改进动力总成控制作用，从而降低燃油消耗。然而，为了学习最优策略而探索多样化经验的需求，往往限制了强化学习技术在车辆控制等安全关键系统中的应用。本文推导了一种指数控制障碍函数（ECBF），并用其过滤基于强化学习的驾驶辅助系统提出的不安全动作。强化学习智能体可自由探索并优化性能目标，同时将不安全动作投影到安全域内距离最近的动作上。奖励的设计方式确保驾驶员的加速度请求得到满足，同时提高燃油经济性且不影响舒适性。通过针对混合动作空间配置的最大后验策略优化（MPO）算法，计算使累积奖励最大化的最优挡位和牵引扭矩控制作用。所提出的安全强化学习方案在跟车场景中进行训练和评估，结果表明，该方案在训练和评估期间均能有效避免碰撞，同时实现驾驶辅助系统预期的燃油经济性提升。

关键词： 强化学习驾驶辅助、安全强化学习、安全过滤、控制屏障函数

arXiv:2301.00904v1 [cs.RO] 3 Jan 2023

1. 引言

可靠、安全且高效的商用车辆对于运输业至关重要，有助于其对环境、经济和道路安全产生积极影响。根据预计到2030年货运需求将增长16%的估算（美国交通统计局，2017年），随着更多车辆上路以满足这一需求，显然有必要也有机会减少排放和燃油消耗。此外，通过预防事故确保安全也至关重要。紧急制动、自适应巡航控制（自适应巡航控制）和车道保持辅助等高级驾驶辅助系统主要用于应对安全问题。更先进的系统还可辅助驾驶员采取环保（节油）的驾驶行为，例如减少制动和加速度（Barkenbus，2010年）、优化自适应巡航控制的速度曲线（Maamria 等，2016年；Nie 和 Farzaneh，2020年），以及优化换挡（Ngo 等，2013年）。

由于典型的商业行驶路线常常需要频繁启停，且要求车辆以各种速度行驶，驾驶员使用自适应巡航控制可能会感到繁琐。Yoon 等人（2020）提出了一种驾驶辅助实现方案，利用雷达信息和运动模型，直接调节动力系统/制动系统的扭矩请求。为此，该方案采用模型预测控制方案，在跟踪驾驶员期望加速度并与前车保持安全距离的同时，优化牵引/制动扭矩/功率。Kerbel 等人（2022）采用类似的驾驶辅助目标，通过一种无模型强化学习方法，利用车辆运行过程中产生的燃油消耗及其他奖励信号，同时学习最优挡位选择和扭矩请求。尽管该研究表明燃油消耗最多可降低 12%，但并未采取措施确保碰撞避免。与模型预测控制等典型最优控制方案不同，后者基于动力学模型设定硬约束；而在强化学习控制器中，学习最优策略需要探索不同的动作和状态，因此通常难以强制施加此类约束。然而，对于车辆控制等安全关键系统而言，无限制探索是不可接受的。本文构建了一种驾驶辅助强化学习智能体，旨在提高燃油效率、实现驾驶员适应性，并纳入确保安全的要素。

为将强化学习智能体的探索恰当地限制在安全集内，研究者提出了不同的方法。在 Li 和 G¨orges（2020）的研究中，采用监督器直接强制执行（覆盖）挡位和发动机转速约束，以控制变速器，但强化学习智能体实际上始终没有学会这些限制。通常会采用奖励塑形方法，对安全违规施加惩罚，以抑制导致约束违规的策略。由于采用奖励塑形的强化学习智能体只有在经历惩罚后才会学到这些惩罚，因此这种方法无法保证安全，尤其是在训练初期。另一种保障安全的方法是将问题建模为约束马尔可夫决策过程（CMDP），其中

为每个状态-动作对分配约束成本，并让强化学习智能体学会将整个时域内折扣后的约束成本控制在某个阈值以下（Altman，1999年）。许多 CMDP 实现随后会联合优化主要性能任务与约束奖励，这就需要在安全性和性能之间进行权衡。在本研究中，我们采用一种称为安全过滤方法的方案，以在一定程度上将这两个目标解耦。该方法使强化学习智能体专注于最大化性能（奖励），同时通过安全层/过滤器将强化学习智能体的输出投影到安全集内。尽管该过滤器通常不会干预强化学习智能体的内部运作，但它会影响性能，因为它往往决定了安全集的范围，以及强化学习智能体随后与受控系统的交互。安全层对所提动作的评估可以基于从离线数据中学习得到的约束（Dalal等人，2018年）和安全指标（Thananjeyan等人，2021年；Srinivasan等人，2020年），也可以使用系统的动力学模型。

在基于动力学模型的安全过滤方法中，控制屏障函数（CBF）能够提供可扩展且计算开销较低的安全滤波器（Li，2021年）。CBF通过强制系统在由连续可微函数 $h(x)$: $R^n \rightarrow R$ 的超水平集定义的不变安全集中运行，从而施加严格的安全约束。强化学习智能体选择的动作会以尽可能少地修改所提动作的方式投影到安全集内（Ames等人，2019年），且任何不安全动作都不会传递给受控系统。可以根据系统动力学设计手工设计的控制障碍函数；例如，在碰撞规避问题中，车辆可用的最大减速度与距离间隔之间的关系就是一个典型案例（Ames等人，2014年；Cheng等人，2019年）。对于本应用中的高相对阶非线性系统，已有研究提出了称为指数障碍函数（ECBF）的专用控制屏障函数（Nguyen和Sreenath，2016年）。

本文推导了一种特定的 ECBF 结构，使其能够与强化学习驾驶辅助智能体协同工作，从而明确考虑惯性效应；这些效应与商用车辆在交通中的安全密切相关。驾驶辅助强化学习智能体的主要性能目标由一个多目标奖励函数规定，该函数旨在平衡驾驶员适应性、燃油经济性和车辆平稳运行。此外，通过为储备功率设置额外奖励，鼓励提升驾驶性能。在这方面，Ngo（2012）仅通过分析不同的标准驾驶循环，来表征给定车速下的加速潜力。本文提出学习储备功率奖励，以便根据驾驶条件和驾驶员习惯定制车辆的响应。

总而言之，本文的贡献如下：1）构建一个驾驶辅助强化学习智能体，用于优化商用车辆的挡位选择和扭矩控制；2）推导一种可与该强化学习智能体配合使用的 ECBF 安全过滤器；3）展示学习功率储备属性的潜力，以便根据驾驶员和驾驶条件进一步定制系统。本文其余部分安排如下：第2节讨论车辆模型和驾驶辅助强化学习智能体。

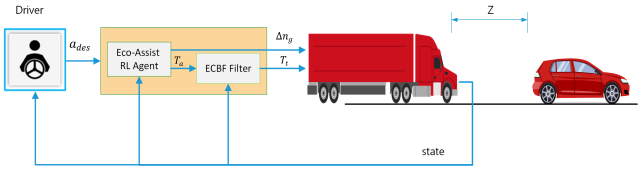


图1. 所提出的基于安全强化学习的节能辅助系统架构

第3节讨论安全滤波器的设计，以及随后将强化学习智能体的输出投影到安全集上的过程。第4节介绍仿真与训练设置，第5节讨论相关结果。最后，第6节总结全文。

2. 车辆环境与强化学习控制器

图1展示了所提出的基于强化学习的驾驶辅助系统示意图，其中包含安全滤波器。本节详细介绍驾驶辅助强化学习智能体的各个计算组件；下一节将讨论 ECBF 安全过滤器。

车辆-驾驶员-环境系统被建模为具有状态 s、动作 a、奖励r 和折扣因子 γ的马尔可夫决策过程（MDP）。状态包含在$s = \{v_l, v_{rel}, a_{des}, a, z, n_g, m_v, \theta, f\}$ 中，分别为自车速度、前车与自车之间的相对速度、驾驶员需求加速度、实际车辆加速度、与前车的间距、变速器挡位、车辆质量、道路坡度，以及一个用于提示前车是否处于自车雷达感知范围内的标志。强化学习控制器通过控制车轮牵引扭矩 T_t 并选择换挡 Δn_g来最大化车辆的性能目标，即动作向量为：$a = \{T_t, \Delta n_g\}$。根据当前挡位的传动比和主减速比，利用车速和车轮牵引扭矩反向计算发动机扭矩和转速。随后，使用燃油消耗率图计算给定发动机扭矩和转速下的燃油消耗。

下文公式 (1) 给出的奖励函数旨在体现驾驶辅助强化学习智能体的性能目标。强化学习智能体的主要目标是满足驾驶员的加速度请求，因此，加速度误差项被赋予了较高权重，w_a。通过权重为 w_f 的燃油消耗率奖励项，鼓励强化学习智能体在满足驾驶员所需加速度的同时，使发动机在燃油效率较高的工况点运行。平顺的扭矩变化的权重为 w_t；为减轻挡位频繁切换及其造成的车辆运行不平顺，还加入了权重为 w_g 的换挡频率惩罚项。请注意，相关奖励信号均按其对应的最大值进行归一化，如下标 max 所示。 $\dot{m}_f$ 和 ΔT_t 分别为燃油消耗率和扭矩变化。

$$r = w_a 0.1^{\frac{|a-a_{des}|}{a_{des,max}}} + w_f 0.1^{\frac{\dot{m}_f}{\dot{m}_{f,max}}} + w_T 0.1^{\frac{|\Delta T_t|}{T_{t,max}}} + w_g 0.1^{\frac{|\Delta n_g|}{n_{g,max}}} + r_{pr}, \quad (1)$$

其中， r_{pr} 表示功率储备奖励项，用于考虑驾驶性能的提升。我们将其定义为： $r_{pr} = P_{res,\ req} - P_{res} w_{pr} 0.1$ 若 $P_{res} < P_{res,\ req}$，否则为 $r_{pr} = w_{pr}$，其中 w_{pr} 为相应的权重。 P_{res} 为实际可用功率，它以发动机转速和发动机扭矩表示为 $P_{res} = (T_{e,\ max}(\omega) - T_e)\omega$；而 $P_{res,\ req}$ 为所需功率储备，我们将在下文讨论。

为了使$P_{res,req}$ 适应不同驾驶条件下不同的加速度需求，Ngo（2012）将加速潜力建模为随车辆速度变化。为此，将车辆速度离散化，并收集多个标准循环中各速度级别对应的加速请求，以拟合累积分布函数。在给定设计置信水平（通常为 90%）下的最大加速度，被视为该速度下所需的加速潜力， $a_{req}(v)$。给定 $a_{req}(v)$后，所需功率储备随后用$P_{res,req} = m_v v a_{req}(v)$建模。在我们的工作中，我们不使用标准驾驶循环中的加速度数据，而是提出使用驾驶员在当前驾驶条件下产生的数据。我们对驾驶员的需求加速度进行连续拟合，以获得 $a_{req}(v)$，从而适应驾驶员需求。我们发现，将 Ngo（2012）提出的 a_{req} 多项式拟合用于驾驶员生成的训练数据时会导致过拟合问题，因此改用一种更易于参数化和学习的逻辑函数：

$$a_{req}(v) = \frac{k_1}{1 + k_2 k_3^{-v}}. \tag{2}$$

接下来，我们简要介绍用于训练驾驶辅助强化学习智能体的框架。状态、控制作用、下一状态及其对应的奖励会持续存储在记忆缓冲区 $\mathcal{R}$ 中。我们采用 Kerbel 等人（2022）提出的演员-评论家架构进行训练，该架构使用一种称为最大后验优化（MPO）的离策略算法（Abdolmaleki等人，2018年；Neunert等人，2020年）。虽然也可以使用其他先进算法，但对于当前问题，我们选择 MPO，因为它样本效率高、对超参数具有鲁棒性，而且易于应用于混合动作空间。该算法从策略评估步骤开始，在此步骤中，评论家网络会估算该策略的状态-动作价值（Q 值）。通过最小化当前 Q 值与估算的目标 Q 值之间的平方损失函数来进行优化， $Q_{target}(s, a)$。本研究采用了 Munos 等人（2016年）所述的 Retrace 算法来计算目标 Q 值，该算法以高效、稳定著称。

对于策略改进（演员网络），MPO 算法采用期望最大化方案。通过从记忆缓冲区中采样，我们构造一个非参数化策略 q，使其最大化 $\mathbb{E}_q[Q_{\theta_k}(s,\ a)]$。

$$\begin{aligned} &\max_q \mathbb{E}_{q(a|s)}[Q_{\theta_k}(s,a)] \\ &s.t.\ \mathbb{E}_{\mu(s)}\left[KL\left(q(a|s)||\pi_{\theta_k}(a|s)\right)\right] < \epsilon, \end{aligned} \tag{3}$$

其中， $\mu(s)$是回放缓冲区中给出的访问分布。随后，拟合出一个新的参数化策略 π_θ，使其符合 q，同时施加 Kullback–Leibler（KL）散度约束，以限制其偏离当前策略的程度。演员网络的参数通过 Adam 求解器（Kingma和Ba，2015）中的基于梯度的优化进行更新。更多关于最大后验优化算法的详细说明，请参见 Neunert 等人（2020）和 Abdolmaleki 等人（2018）。本应用的更多实现细节也可参见 Kerbel 等人（2022）中我们直接强化学习实现的相关内容。如上所述，其他最先进的强化学习训练算法也可用于驾驶辅助强化学习智能体，且这与下文讨论的安全滤波器无关。

3. 指数控制屏障函数安全滤波器

本节将推导适用于我们应用的指数控制屏障函数滤波器。首先简要回顾控制屏障函数和指数控制屏障函数的定义。有关这些主题的更详细讨论，请参阅Nguyen和Sreenath（2016年）。考虑以下非线性控制仿射系统：

$$\dot{x} = f(x) + g(x)\,u, \tag{4}$$

其中， f 和 g 是局部利普希茨连续的， $x \in \mathcal{R}^n$ 是系统的状态， $u \in \mathcal{R}^m$ 是控制输入。假设安全集定义为 $\mathcal{C} = \{x \in \mathcal{R}^n | h(x) \geq 0\}$，其中 h： $\mathcal{R}^n \to \mathcal{R}$是连续可微函数。若存在一个扩展类 κ_∞ 函数 α，使得对所有 $x \in Int(\mathcal{C}) = \{x \in \mathcal{R}^n :\ h(x) > 0\}$ 均成立，则 h 是控制屏障函数：

$$\sup_{u \in U} \left[L_f h(x) + L_g h(x)\ u\right] \geq -\alpha(h(x)). \tag{5}$$

h 是控制屏障函数这一事实确保安全集 $\mathcal{C}$ 是正向不变的，因此我们能够保证安全。指数控制障碍函数通过对相对阶为 r 的非线性系统进行输入-输出（IO）线性化来生成控制屏障函数。如Nguyen和Sreenath（2016）所述，新的虚拟线性系统（经过IO线性化后）的状态变量为$\eta_b := [h(x)\ \dot{h}(x)\ \cdots,\ h^r(x)]^T$，输入为 μ，输出为 $h(x)$：

$$\begin{aligned} \dot{\eta}_b &= F\eta_b(x) + G\mu, \\ h(x) &= C\eta_b, \end{aligned} \tag{6}$$

其中， F 和 G 是表示积分器链的矩阵， $C = [1,\ 0,\ \cdots,\ 0]$。

虚拟线性系统 μ 的控制作用是控制输出的 r^{th} 阶导数； $\mu = L_f^r h(x) + L_g L_f^{r-1} h(x) u$。当 μ 由增益为 K_α的状态反馈控制设为 $\mu = -K_\alpha \eta_b$时，控制输出随时间的变化为 $h(x(t)) = Ce^{(F - GK_\alpha)\ t}\eta_b(x_0)$。对于初始条件 $h(x_0) > 0$，通过施加 $\mu > -K_\alpha \eta_b(x)$，可以保证 $h(x(t)) \geq Ce^{(F - GK_\alpha)\ t}\eta_b(x_0)$。这一关系引出了指数控制障碍函数的定义。考虑动态系统（ $\mathcal{C} = \{x \in \mathcal{R}^n | h\ x\ \geq$ ）和 $0\}\ h\ x$ $K_\alpha \in \mathcal{R}^{r \times 1}$ 集合（），若存在，则（）是指数控制障碍函数

$$sup\left[L_f^r h(x) + L_g L_f^{r-1} h(x)\, u\right] \geq -K_\alpha \eta_b(x),\ \forall\ x \in \mathcal{C}. \tag{7}$$

如果 K_α 使闭环系统矩阵 $F - GK_\alpha$比赫尔维茨条件更强，且选择合适的初始条件，则 $\mu \geq -K_\alpha \eta_b(x)$可保证 $h(x)$是指数控制障碍函数。随后可使用线性反馈控制的极点配置策略来设计该指数控制障碍函数。

我们以间距 z、自车速度 v_h 和前车速度 v_l 作为状态变量，构建避碰模型。模型如下：

$$\dot{z} = v_l - v_h, \tag{8a}$$

$$\dot{v_l} = a_l, \tag{8b}$$

$$\dot{v}_h = \frac{T_t}{r_w m_v} - \frac{F_r\left(v_h, m_v, \theta\right)}{m_v}, \tag{8c}$$

$$F_r = \frac{\rho A c_d v_h^2}{2} + m_v g f \cos\theta + m_v g \sin\theta, \tag{9}$$

其中， F_r 是总阻力，包括重力、滚动摩擦阻力和空气阻力。T_t， c_d， f， θ， m_v， ρ， A_v， r_w， a_l 分别表示车轮上的牵引扭矩、空气阻力系数、滚动阻力系数、道路坡度、车辆质量、空气密度、车辆迎风面积、车轮半径和前车加速度。

由于具有仿射动力学，上述状态空间表示也可以在写成式（4）时分解为未驱动动力学 f（x）和受驱动动力学 g（x）部分。对于最小车间距目标 z_0，一个自然的选择是 h（x）$=$ $z - z_0$。随后采用输入输出线性化，将非线性动力学转换为式（6）中的虚拟线性系统；结合上文的相关讨论，并取反馈增益 $K_\alpha = [k_{\alpha 1}, k_{\alpha 2}]$，可得：

$$\dot{h}(x) = v_l - v_h, \tag{10}$$

$$\mu = \ddot{h}\left(x\right) = \frac{F_r\left(v_h, m_v, \theta\right)}{m_v} + a_l - \frac{T_t}{m_v r_w}, \tag{11}$$

$$\begin{aligned} -K_\alpha \eta_b\left(x\right) &= -k_{\alpha 1} h(x) - k_{\alpha 2}\dot{h}(x) \\ &= -k_{\alpha 1}\left(z - z_0\right) - k_{\alpha 2}\left(v_l - v_h\right). \end{aligned} \tag{12}$$

利用这些结果和式（7），即可得到指数控制屏障函数滤波器。强化学习智能体提出的牵引扭矩动作随后会先通过该安全滤波器，再发送至车辆环境。如图1所示，指数控制屏障函数滤波器通过将强化学习智能体提出的动作 T_a（s）投影到安全控制牵引扭矩 T_t 上来实现安全约束，同时以尽可能小的改动完成投影，具体由下方的二次规划问题给出。

$$\begin{aligned} T_t^* = &\arg\min_{T_t} \frac{1}{2}\left\|T_t - T_a(s)\right\|^2 \\ \text{s.t.} \quad & a_l + \frac{F_r\left(v_h, m_v, \theta\right)}{m_v} - \frac{T_t}{m_v r_w} \geq -k_{\alpha 1}\left(z - z_0\right) \\ & - k_{\alpha 2}\left(v_l - v_h\right) \end{aligned} \tag{13}$$

4. 仿真与训练设置

通过对搭载 10 速自动手动变速器（AMT）的中型卡车进行仿真，展示带有 ECBF 安全过滤器的强化学习驾驶辅助智能体的工作原理和性能。策略网络和价值网络均由深度神经网络表示，包含三个隐藏层，每层有 256 个节点。强化学习控制器在这样一种场景中进行训练：由不完美驾驶员驾驶的卡车在联邦测试程序（FTP-75）驾驶循环（Barlow等人，2009）下跟随前车。考虑到商用车辆必须应对大幅的载荷波动，智能体的训练重量范围为 5 至 10 吨。为帮助涵盖不同的驾驶体验，训练数据通过以下方式进行随机化：在前车的速度曲线上添加噪声，随机设置初始间距和道路坡度，并调整驾驶员模型的参数；本研究采用的驾驶员模型为智能

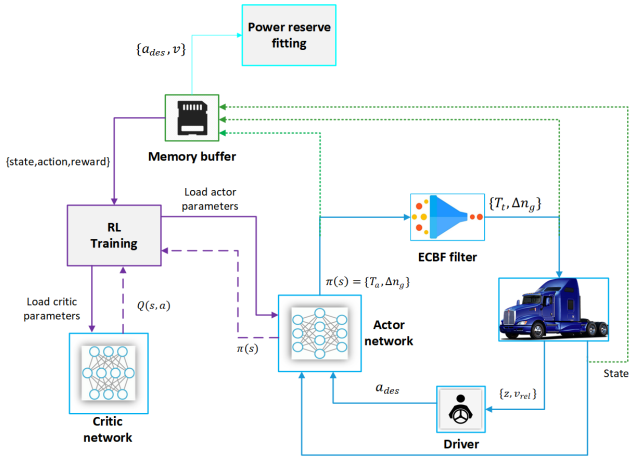


图2. 强化学习训练设置

驾驶员模型（IDM）（Treiber等人，2000），其方程如下。IDM 的期望最小间隙按式（15）计算，其中，当本车接近前车时，接近项 z_{app} （16）起主导作用。我们仅在车距小于某个阈值时添加 z_{app}，超过该阈值后则忽略这一项，以模拟分心驾驶员提出不安全操作请求的情形。

$$a = a_{max}\left(1 - \left(\frac{v}{v_0}\right)^4 - \left(\frac{z^*\left(v, v_{rel}\right)}{z}\right)^2\right), \tag{14}$$

$$z^*\left(v, v_{rel}\right) = z_0 + \tau v - z_{app}, \tag{15}$$

$$z_{app} = \frac{v_{rel} v}{2\sqrt{a_{max} b}}, \tag{16}$$

其中， a_{max} 为最大加速度， b 为舒适的减速度水平， τ 为时距。

在每个仿真步中，如图2所示，IDM驾驶员根据与前车的距离间隔和车速请求加速度。为满足驾驶员需求，演员网络输出 $\{T_a, \Delta n_g\}$，所提动作由ECBF安全层进行筛选。随后，在车辆环境中执行安全动作并获得奖励。如前所述，奖励旨在满足驾驶员的加速度需求，同时促进燃油经济性、驾驶性能和车辆平稳运行。因此，奖励目标的权重设为 $[w_a = 0.65, w_f = 0.2, w_T$ =0.05， $w_g = 0.05$， $w_{pr} = 0.05]$。与此同时，驾驶员的加速度需求和车速用于表征功率储备。对驾驶员的最大加速度请求（90%置信水平）进行逻辑函数拟合，从而按照第2节所述持续调整功率储备项。

对于上述车辆以及表1中给出的参数，我们设计了一个增益向量为 $K_\alpha = [0.8, 2]$ 的 ECBF安全层。我们发现，对于给定的车辆，该过滤器能够有效地将不安全动作投影为安全动作，并且整个训练过程中没有发生碰撞。我们观察到，随着训练的推进，强化学习智能体逐渐学会控制车辆的加速度，以满足驾驶员的要求。随着训练的进行，英里/加仑也有所提升，这表明强化学习智能体成功学会了

以节能方式实现加速度跟踪目标。出于篇幅考虑，我们省略了训练进程的详细信息，转而在下一节展示比较评估。

我们用于驾驶辅助强化学习智能体仿真和训练的参数如表1所示。

表1。车辆环境和强化学习超参数设置

车辆参数		MPO超参数	
Mass	5 - 10 吨	策略网络、价值网络学习率	$10^{-5}, 10^{-5}$
A_u	$7.71m^2$	对偶约束	0.1
C_d	0.08	Retrace步数	15
r_w	0.498	KL约束 ϵ_μ, ϵ_σ, ϵ_d	0。1, 0。001, 0。1
f	0.015	α_d, α_c	10
z_{sr}	350	γ	0.99

作为基准，我们还考虑并模拟了不包含安全强化学习辅助智能体的相同驾驶场景（仅 IDM）。基准动力总成控制会产生牵引扭矩，以抵消阻力并满足 IDM驾驶员要求的加速度。对于挡位决策，我们根据 Yoon 等人（2020）所述的方案，选择燃油消耗率最低的最佳挡位。该方案基于模型，并且完全掌握发动机燃油消耗图。请注意，我们的强化学习智能体并不了解发动机燃油消耗图或任何建模动力学。

5. 评估结果与讨论

在 ARTEMIS 城市驾驶循环（Barlow等人，2009）下跟随前车的过程中以及训练结束后，对 RL-ECBF辅助的安全性能进行了评估；该驾驶循环不同于训练所采用的 FTP 循环。图3展示了 RL-ECBF辅助如何应对训练中的最坏情况：RL探索和分心驾驶员都是产生不安全动作的来源。分心驾驶员由 IDM 建模，其 τ 仅考虑距离间隔小于 $50m$时的接近项（z_{app}）。该驾驶员缩短了初始 $350m$ 间距，并与前车发生碰撞（图3.A中的红色星标）。然而，引入 RL-ECBF辅助后，由于显著的 ECBF 投影（图3.D），且 RL 开始意识到安全边界，没有不安全动作被发送至车辆环境。

我们还考察了强化学习—指数控制屏障函数辅助系统在配合谨慎驾驶员时，满足驾驶员需求和提升燃油效率的表现。为评估这一方面，我们通过启用接近项（z_{app}）来模拟一位相对谨慎（驾驶表现良好）的驾驶员，该驾驶员采用 IDM驾驶员模型，且车距小于$100m$。图 4 和表 2 展示了良好 IDM 驾驶员单独驾驶（采用基于模型的挡位和扭矩控制）与同一驾驶员在训练良好的强化学习智能体及 ECBF 安全过滤器辅助下的性能比较。对于 IDM 驾驶员单独驾驶的情况，驾驶员需求与实际车辆加速度之间的均方根误差为 0.38；引入辅助系统后，该数值降至 0.17。除了提升驾驶性能，强化学习智能体还倾向于使用较高挡位，从而提高英里/加仑

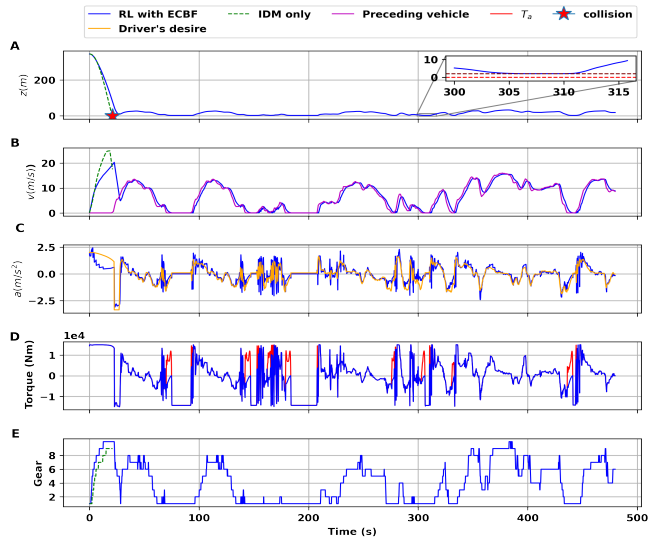


图 3. 对短视 IDM驾驶员进行若干轮训练后，评估带 ECBF 滤波器的强化学习辅助系统。(A) 中的红色星号表示未使用强化学习智能体（仅 IDM）时发生碰撞的情况。

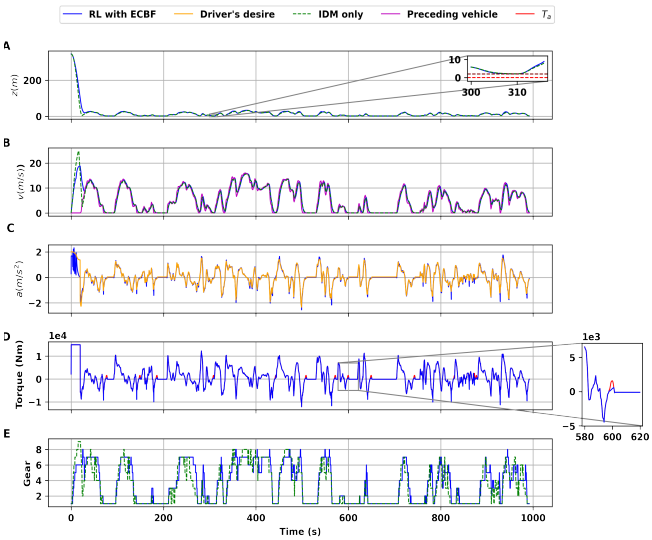


图4. 对谨慎驾驶员使用指数控制障碍函数滤波器的训练充分的强化学习辅助系统进行评估

相比基准（仅采用 IDM 和基于模型的动力总成控制），性能提升了 6.34%。强化学习智能体最终学会将运行状态主要限制在指数控制屏障函数滤波器定义的安全集内，因为对不安全动作进行投影通常会导致次优行为（奖励较低）。图 4.D 展示了这一点：指数控制屏障函数的投影相对少见、幅度较小，且仅出现在近距离快速接近时。

表2. 使用和不使用RL-ECBF辅助的谨慎驾驶员性能比较

	不带辅助的IDM RL-ECBF辅助	带RL-ECBF的IDM 辅助
MPG	6.875(–)	7.31(6.34%)
$a_{rms}(m)$	0.38	0.17
$Z_{mean}(m)$	15.5	16.7
$Z_{min}(m)$	2	2

6. 结论

本文提出并验证了一种驾驶辅助强化学习智能体，可协助驾驶员提升燃油经济性和驾驶性能。训练和实际运行期间，该智能体均通过指数控制障碍函数（ECBF）过滤不安全动作，从而确保安全性。RL-ECBF辅助系统经过训练，可最大化多目标奖励结构，在满足驾驶员加速度需求、燃油经济性、平稳运行和功率储备等目标之间取得平衡。在训练强化学习智能体的过程中，针对特定驾驶员持续调整加速请求曲线，以确保其拥有足够的加速潜力。在不同于训练所用驾驶循环的测试中，评估结果表明，RL-ECBF辅助系统在确保安全的同时，成功提升了燃油经济性和驾驶性能；即使考虑到分心驾驶员在没有驾驶辅助系统介入时会引发碰撞的情况，结果依然如此。

未来的工作中，我们计划使用随机化交通数据和仿真来训练并评估所提出的强化学习-指数控制障碍函数智能体。此外，还需要考虑此处概述的基于模型的指数控制障碍函数投影方法中固有的不确定性。

参考文献

Abdolmaleki, A., Springenberg, J.T., Tassa, Y., Munos, R., Heess, N., and Riedmiller, M.A.（2018年）。最大后验策略优化。*CoRR*，abs/1806.06920。网址：http://arxiv.org/abs/1806.06920。Altman, E.（1999年）。受约束马尔可夫决策过程。Ames, A.D., Coogan, S., Egerstedt, M., Notomista, G., Sreenath, K., and Tabuada, P.（2019年）。控制屏障函数：理论与应用。*2019*年欧洲控制会议第*18*届会议，*ECC 2019*，3420–3431。Ames, A.D., Grizzle, J.W., and Tabuada, P.（2014年）。基于控制屏障函数的二次规划及其在自适应巡航控制中的应用。*IEEE* 决策与控制会议论文集，2015年2月，6271–6278。Barkenbus, J.（2010年）。生态驾驶：一项被忽视的气候变化倡议。《能源政策》，38，762–769。Barlow, T.J., Latham, S., Mccrae, I.S., and Boulter, P.G.（2009年）。道路车辆排放测量用驾驶循环参考手册。美国交通统计局（2017年）。货运分析框架，第5版。Cheng, R., Orosz, G., Murray, R.M., and Burdick, J.W.（2019年）。通过屏障函数实现面向安全关键型连续控制任务的端到端安全强化学习。第*33*届*AAAI* 人工智能会议，*AAAI 2019*，3387–3395。Dalal, G., Dvijotham, K., Vecerik, M., Hester, T., Paduraru, C., and Tassa, Y.（2018年）。连续动作空间中的安全探索。Kerbel, L., Ayalew, B., Ivanco, A., and Loiselle, K.（2022年）。基于深度强化学习的驾驶辅助生态驾驶与变速器控制。Kingma, D.P. and Ba, J.（2015年）。Adam：一种随机优化方法。国际学习表征会议。Li, G. and G¨orges, D.（2020年）。基于强化学习的阶梯式变速器车辆生态自适应巡航控制。*IEEE* 智能交通系统汇刊，21(11)，4895–4905。doi：10.1109/TITS.2019.2947756。Li, Z.（2021年）。安全方法比较：控制屏障函数与可达性分析。*arXiv* 预印本 *arXiv:2106.13176*。Maamria, D., Gillet, K., Colin, G., Chamaillard, Y., and Nouillant, C.（2016年）。论动态规划在电动汽车经济驾驶循环计算中的应用。1288–1293。doi:10.1109/CCA.2016.7587984。Munos, R., Stepleton, T., Harutyunyan, A., and Bellemare, M.G.（2016年）。安全高效的离策略强化学习。载于第*30*届神经信息处理系统国际会议论文集，NIPS'16，1054–1062。Curran Associates公司，美国纽约州雷德胡克。Neunert, M., Abdolmaleki, A., Wulfmeier, M., Lampe, T., Springenberg, J.T., Hafner, R., Romano, F., Buchli, J., Heess, N., and Riedmiller, M.A.（2020年）。用于机器人混合控制的连续-离散强化学习。*CoRR*，abs/2001.00449。网址：http://arxiv.org/abs/2001.00449。Ngo, V.D.（2012年）。汽车变速器换挡策略。2012年。doi:10.6100/IR735458。Ngo, V.D., Navarrete, J.A.C., Hofman, T., Steinbuch, M., and Serrarens, A.（2013年）。兼顾燃油经济性与驾驶性能的最优换挡策略。英国机械工程师学会会刊*D*辑：汽车工程学报，227(10)，1398–1413。doi：10.1177/0954407013491240。Nguyen, Q. and Sreenath, K.（2016年）。用于实施高相对阶安全关键约束的指数型控制屏障函数。美国控制会议论文集，2016年7月(3)，322–328。Nie, Z. and Farzaneh, H.（2020年）。基于模型预测控制算法的生态驾驶自适应巡航控制。*AppliedSciences*，10，5271。doi：10.3390/app10155271。Srinivasan, K., Eysenbach, B., Ha, S., Tan, J., and Finn, C.（2020年）。学习如何保持安全：结合安全评论家的深度强化学习。1–16。Thananjeyan, B., Balakrishna, A., Nair, S., Luo, M., Srinivasan, K., Hwang, M., Gonzalez, J.E., Ibarz, J., Finn, C., and Goldberg, K.（2021年）。恢复强化学习：利用学习得到的恢复区域实现安全强化学习。*IEEE*机器人学与自动化快报，6(3)。Treiber, M., Hennecke, A., and Helbing, D.（2000年）。基于经验观测和微观仿真的拥堵交通状态。*Physical Review E*，62，1805–1824。doi:10.1103/PhysRevE.62.1805。Yoon, D.D., Ayalew, B., Ivanco, A., and Loiselle, K.（2020年）。用于重型车辆加装型驾驶辅助生态驾驶系统的预测性动能管理。*IET*智能交通系统，14(13)，1824–1834。