基于强化学习的自适应巡航控制安全过滤

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，邮编 *29607*（*hhailem*、*beshah*、*lsutto2*）*@clemson.edu*。

**艾里逊变速箱公司，美国印第安纳州印第安纳波利斯 *One Allison Way*，邮编 *46222*（*andrej.ivanco*、*keith.loiselle*）*@allisontransmission.com*

摘要： 能够学习并适应道路、交通状况及车辆状况的强化学习（RL）自适应巡航控制系统（ACC），有望提升车辆能源效率并改善交通流。然而，将强化学习应用于自适应巡航控制等安全关键系统，需要强有力的安全保证；而学习智能体从根本上需要进行探索，因此很难实现这些保证。本文推导出控制屏障函数，并将其用作安全过滤器，使基于强化学习的自适应巡航控制器能够在无碰撞安全集内自由探索。具体而言，我们针对高相对阶非线性系统推导控制屏障函数，以考虑与商用车辆相关的惯性效应。我们还概述了一种利用这些屏障函数处理执行器饱和的算法。尽管任何强化学习算法都可与这些过滤器结合，用作性能型自适应巡航控制器，但我们采用了最大后验策略优化（MPO）算法，并使用混合动作空间来学习燃油最优挡位选择和扭矩控制策略。我们将安全过滤强化学习方法与奖励塑形强化学习方法进行了比较；后者只有经过充分训练后才能学会避免碰撞。在不同驾驶循环上的评估表明，与基准自适应巡航控制算法相比，所提出的方法显著提升了燃油经济性。

关键词： 自适应巡航控制、安全强化学习、安全过滤、控制屏障函数

arXiv:2301.00884v1 [eess.SY] 2 Jan 2023

1. 引言

自适应巡航控制（ACC）系统是现代车辆中日益普及的驾驶辅助系统之一。ACC 系统利用雷达、计算机视觉或激光雷达感知车辆周围环境，并据此做出控制决策。当其他车辆或物体不在感知范围内时，ACC 会补偿道路坡度、摩擦力和空气阻力，以维持驾驶员设定的车速。当前方有其他车辆或物体时，ACC 会做出决策以避免碰撞，并尽可能贴近前车行驶，从而避免车辆加塞。研究表明，ACC 能够减轻驾驶员的工作负担，并使交通流更加高效、安全（Marsden 等人，2001年；Lang 等人，2014年）。

有效的自适应巡航控制系统应兼顾道路交通状况、车辆性能和驾驶员的期望速度。目前可用的基于 PID 的自适应巡航控制系统（Canale 和 Malan，2003年；Chamraz 和 Balogh，2018年）以及已提出的基于模型预测控制的方法（Naus 等人，2008年；Yang 等人，2021年）通常会针对各种运行环境进行调校，以平衡这些因素。尽管可以采用“自适应”或增益调度版本（Radke 和 Isermann，1987年），但这些方法的固定结构限制了车辆在整个使用寿命期间的全面适应能力。此外，基于模型预测控制的自适应巡航控制还必须找到可靠的方法，预测未来时域内前车的运动。另一方面，数据驱动强化学习方法（Li 和 G¨orges，2020年）无需预先设定控制架构，即可持续适应交通状况、道路状况和车辆状况。在本研究中，我们探讨强化学习自适应巡航控制在商用车辆中的应用。此外，传统自适应巡航控制主要关注速度跟踪和保持安全车距这两项任务，而我们考虑通过强化学习自适应巡航控制（简称强化学习自适应巡航控制）明确优化燃油经济性，具体借助挡位选择和扭矩控制策略实现。

尽管自适应能力和性能提升具有潜在优势，强化学习自适应巡航控制仍面临严峻的安全挑战。这些挑战源于强化学习算法需要通过探索来学习最优策略。强化学习在经历给定的状态-动作对后，才能了解其优劣；但对于车辆控制等应用，即使在强化学习算法的（道路上）训练期间，在不安全域中探索也是不可接受的。不过，得益于安全强化学习的最新进展，人们提出了不同的方法，以鼓励或限制探索，使其仅在安全域内进行。下面简要介绍其中几种方法。奖励塑形方法会在违反约束时，对性能目标函数施加较大的惩罚。另一方面，约束马尔可夫决策过程（CMDP）方法为每个状态-动作对分配安全约束成本，并将轨迹的总安全约束成本限制在某个阈值以下（Altman，1999）。

奖励塑形和CMDP方法在性能控制器本身中实现，以鼓励遵守安全约束，但无法保证安全。另一类方法则采用施加硬性约束的安全过滤器。这些方法将以性能为导向的强化学习控制器与安全过滤器分离：前者唯一的目标是优化系统的性能目标函数，后者则将性能控制器提出的不安全动作投影到安全集中。安全过滤器利用系统的动力学模型确定给定状态-动作对的安全条件，或利用离线数据学习约束（Dalal 等，2018）和安全指标（Thananjeyan 等，2021；Srinivasan 等，2020）。本文采用基于动力学模型的安全保证来构建安全集，使强化学习性能控制器能够在安全边界内自由探索。随着训练的推进，强化学习性能控制器最终会学会安全边界，并不再提出不安全动作（Thananjeyan 等，2021）。请注意，尽管安全过滤器不会干预控制器的内部运行机制，但它会通过规定性能控制器可运行的范围，影响控制性能。

在基于模型的安全过滤器设计方法中，控制屏障函数（控制屏障函数）具有计算量小、可扩展性强的优点（Li，2021）。控制屏障函数通过使控制器在由连续可微函数的超水平集定义的不变安全集内运行，来保证安全性$h(x): \mathbb{R}^n \to \mathbb{R}$。性能控制器选取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，2019），并且不会将不安全动作传递给受控系统。指定控制屏障函数可以采用多种方法，各有优缺点。一种直观的方法是根据系统动力学及其相关的动作边界，手工构造控制屏障函数（Xu 等，2018；Ames 等，2014；Cheng 等，2019）。例如，在碰撞规避问题中，可以根据系统能够施加的最大减速度推导控制屏障函数，以缩短距离间隔。在条件允许时，还希望逐步扩大安全集，以获得最大的安全域；目前，对于多项式对象动力学，可以通过平方和（平方和）规划和多项式控制屏障函数实现这一目标（Chamraz 和 Balogh，2018）。另一种适用于相对阶较高的非线性动力系统（例如涉及惯性效应的系统）的方法，是采用指数控制屏障函数（ECBF）（Nguyen 和 Sreenath，2016）。在本研究中，我们推导指数控制屏障函数，使其与我们的强化学习自适应巡航控制器协同用作安全过滤器，从而明确考虑惯性效应；对于载荷变化较大的商用车辆而言，这一点十分重要。

因此，本文的主要贡献是推导并展示了一种基于控制屏障函数的安全强化学习自适应巡航控制方法，适用于优化燃油经济性的商用车辆。虽然我们推导了用于安全认证的指数控制屏障函数，但需要指出的是，直接使用指数控制屏障函数（或一般的控制屏障函数）时，通常假定动作不受限制；其原始形式可能会要求车辆动力系统执行无法实现的动作。因此，我们提出了一种方法，可在车辆动作限制范围内，针对给定参数的指数控制屏障函数提供安全保证。我们提出的强化学习自适应巡航控制在优化燃油消耗的目标下协调牵引扭矩控制与挡位决策。采用安全证书增强的强化学习自适应巡航控制在不同驾驶循环上进行训练和评估，并将车辆性能与采用奖励塑形方法实现安全强化学习的强化学习自适应巡航控制，以及传统的基于PID的自适应巡航控制进行比较。

本文其余部分安排如下。第2节介绍我们如何推导出用于自适应巡航控制安全过滤器的指数控制屏障函数，并详细说明如何在其中处理执行器约束。第3节介绍我们的性能强化学习自适应巡航控制算法的具体细节。第4节讨论结果并进行分析，第5节总结全文。

2. 自适应巡航控制的安全过滤器

我们简要回顾控制屏障函数的定义如下。详情见Hsu等人（2015）。考虑如下非线性控制仿射系统：

$$\dot{x} = f(x) + g(x)\,u,. \tag{1}$$

其中，f 和 g 是局部利普希茨连续的，$x \in \mathcal{R}^n$ 是系统状态，$u \in \mathcal{R}^m$ 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。那么，如果存在一个扩展类 κ_∞ 函数 α，使得对所有 $x \in Int(\mathcal{C}) = \{x \in \mathcal{R}^n$ 均有 $h(x) > 0\}$，则 h 是一个控制屏障函数（CBF）：

$$\sup_{u \in U} \left[L_f h(x) + L_g h(x)\,u\right] \geq -\alpha(h(x)). \tag{2}$$

对于高相对阶非线性仿射系统，可以使用反馈线性化来构造指数控制障碍函数（ECBF），具体方法详见 Nguyen和Sreenath（2016）。该方法通过将高相对阶非线性系统转换为一个带有新状态变量的虚拟线性系统（输入输出线性化）来实现：$= [h(x)$ $h(x)$ ，$h^r(x)]^T$，输入 μ 和输出 $h(x)$：

$$\begin{aligned} \dot{\eta}_b &= F\eta_b(x) + G\mu, \\ h(x) &= C\eta_b \end{aligned} \tag{3}$$

其中，F 和 G 是表示积分器链的矩阵，且 $C = [1, 0, \cdots, 0]$。可以为变换后的系统设计状态反馈控制器：$\mu = -K_\alpha \eta_b$，其中增益向量 K_α 应使 $F - GK_\alpha$满足赫尔维茨稳定。对于相对阶为 r的系统，μ也是输出 $h(x)$的 r^{th}阶导数，$\mu = L_f^r h(x) + L_g L_f^{r-1} h(x)u$。如果存在一个状态反馈增益 K_α，使得对于所有状态均有 $\mu \geq -K_\alpha \eta_b(x)$，那么可以证明 $h(x)$是一个指数控制障碍函数（参见 Nguyen和 Sreenath（2016））。

本问题中的自适应巡航控制部分以间距 z、本车速度 v_h 和前车速度 v_l 为状态变量。相应的状态方程为：

$$\dot{z} = v_l - v_h \tag{4a}$$

$$\dot{v_l} = a_l \tag{4b}$$

$$\dot{v_h} = \frac{T_t}{r_w m_v} - \frac{F_r(v_h, m_v, \theta)}{m_v} \tag{4c}$$

$$F_r = \frac{\rho A c_d v_h^2}{2} + m_v g f \cos\theta + m_v g \sin\theta \tag{5}$$

其中， F_r 为总阻力，包括重力、滚动阻力和空气阻力； T_t 为车轮处的牵引扭矩。参数 c_d、 f、 θ、 m_v、 ρ、 A_v、 r_w 和 a_l 分别为空气阻力系数、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。

我们发现，上述模型可以轻松写成控制仿射形式 (1)。给定碰撞安全目标，我们希望间距 z 始终大于指定的最小车间距离 z_0。为此，我们将控制屏障函数 (CBF) 定义为输出 $h(x) = z - z_0$。考虑到控制执行量为牵引扭矩 T_t，我们得到一个相对阶为二的控制仿射系统。从物理角度来看，安全目标针对的是位置，而牵引扭矩直接调节加速度，因此必须考虑惯性效应。将系统通过输入输出线性化变换为形式 (3)，可得：

$$\dot{h}(x) = v_l - v_h, \tag{6}$$

$$\mu = \ddot{h}(x) = \frac{F_r(v_h, m_v, \theta)}{m_v} + a_l - \frac{T_t}{m_v r_w}, \tag{7}$$

$$-K_\alpha \eta_b(x) = -k_{\alpha 1}(z - z_0) - k_{\alpha 2}(v_l - v_h) \tag{8}$$

现在，我们考虑控制输入 μ 的执行限制，计算其在牵引扭矩限值（T_{min} 和 T_{max}）下的一些边界。给定前车加速度（a_l）和本车速度（v_h），μ 的可行边界为

$$\mu_{T_{min/max}} = a_l + \frac{F_r(v_h, \theta, m_v)}{m_v} - \frac{T_{min/max}}{m_v r_w} \tag{9}$$

对于给定的增益向量 $K_\alpha = [k_{\alpha 1}, k_{\alpha 2}]$，如果所提出的状态反馈控制 $-k_{\alpha 1}(z - z_0) - k_{\alpha 2}(v_l - v_h)$ 位于虚拟线性系统的动作边界 $[\mu_{Tmax}, \mu_{Tmin}]$ 内，则指数控制屏障函数 (ECBF) 可保证安全。然而，在一般应用场景中，这一边界可能无法满足。尽管如此，如果选择 K_α，使极点在 s 平面上足够靠左，上述指数控制屏障函数仍可能约束安全集。通过考察最坏情况下控制屏障函数控制项 $h(x)$ 的演化，可以为此类极点选取提供安全保证：将线性虚拟模型以极端可能的 $\eta_{0,xrm}$ 初始化，然后施加可能的限幅扭矩动作。给定最小间隔距离目标和最大下坡坡度，这等效于施加性能强化学习自适应巡航控制智能体可能输出的最大牵引扭矩，同时将最大载荷下的本车模型以可能的最大速度初始化，并令前车处于最大减速度状态。这些极端条件根据方程 (9) 得到可行的 μ 边界 $\mu_{Tmin-xrm}$ 和 $\mu_{Tmax-xrm}$。

为捕捉极端条件下 $h(x)$ 的演变过程，将一次仿真展开离散为时间步长为 Δt 的序列，并使动作 μ（受 $\mu_{Tmin-xrm}$ 和 $\mu_{Tmax-xrm}$ 限幅）保持分段常数。算法 1 展示了如何通过对虚拟系统 (3) 进行积分来实现这一过程。如果该仿真中的 $h(x)$ 在无穷远处（或经过某个有限时间后）为正，则所选的 K_α 可保证安全性。否则，需要更改 K_α，直到满足这一条件。

算法 1 一种在虚拟线性系统上强制执行系统边界的算法

$\eta \leftarrow \eta_0$
$\mu \leftarrow \mu_0$
当 $t \le t$ **时** ∞ **do**
  $t \leftarrow t + \Delta t$
  if $\mu < \mu_{Tmax-xrm}$，**则**
    $\mu \leftarrow \mu_{Tmax-xrm}$
  否则，如果 $\mu > \mu_{Tmin-xrm}$，**则**
    $\mu \leftarrow \mu_{Tmin-xrm}$
  结束 if
  $h(x(t)) \leftarrow C(e^{F\Delta t}\eta_0 + e^{F\Delta t}\int_0^{\Delta t} e^{-F\tau} G\mu d(\tau))$
  $\mu \leftarrow -k_{\alpha 1} h(x) - k_{\alpha 2} h(x)$
  $\eta_0 \leftarrow \begin{bmatrix} h(x) \\ \dot{h} \end{bmatrix}$
结束循环

选定合适的增益向量 K_α 后，指数控制障碍函数安全约束通过将强化学习控制器的演员网络 $T_a(s)$（见下一节）输出所提出的动作，投影到控制牵引转矩 T_t 上，以尽可能减少对该动作的改动，从而确保安全。具体做法是构建并求解以下二次规划：

$$\begin{aligned} T_t^* = & \arg\min_{T_t} \frac{1}{2} \|T_t - T_a(s)\|^2 \\ \text{s.t.} \quad & a_l + \frac{F_r(v_h, m_v, \theta)}{m_v} - \frac{T_t}{m_v r_w} \ge -k_{\alpha 1}(z - z_0) \\ & - k_{\alpha 2}(v_l - v_h) \end{aligned} \tag{10}$$

3. 车辆环境与强化学习自适应巡航控制

动力总成控制器被建模为一个马尔可夫决策过程（MDP），由状态 s、动作 a、奖励函数 $r(s, a)$ 和折扣因子 γ 组成。动作选择的概率由策略 $\pi(a|s, \boldsymbol{\theta})$ 决定，其中 $\boldsymbol{\theta}$ 表示用于近似该策略的深度神经网络参数。本车速度 v_l、前车与本车之间的相对速度 v_{rel}、两车之间的间距 z、挡位 n_g、车辆质量 m_v、道路坡度 θ、驾驶员需求速度 v_{set} 以及用于指示车辆是否处于自适应巡航控制传感器范围内的标志 f，共同构成强化学习智能体的状态 $s = \{v_l, v_{rel}, z, n_g, m_v, \theta, v_{set}, f\}$。强化学习性能控制器旨在同时执行牵引扭矩 T_a 控制和换挡选择 Δn_g，即 $a = \{T_a, \Delta n_g\}$。如图1所示，所提出的 T_a 经由ECBF安全层滤波，得到安全的牵引扭矩需求 T_t（10）。随后，利用所选挡位的变速器传动比和主减速器传动比，计算产生该车轮牵引扭矩的发动机扭矩和发动机转速，并从燃油特性图中读取相应的燃油消耗率。请注意，尽管强化学习控制器的动作是 T_a 和 Δn_g，ECBF安全滤波器在安全约束中并不使用 Δn_g。然而，鉴于挡位选择对燃油经济性和驾驶员舒适性至关重要，它是强化学习性能控制器不可或缺的一部分。

经过滤波的牵引扭矩 T_t 和换挡 Δn_g 动作在车辆环境中执行，并且

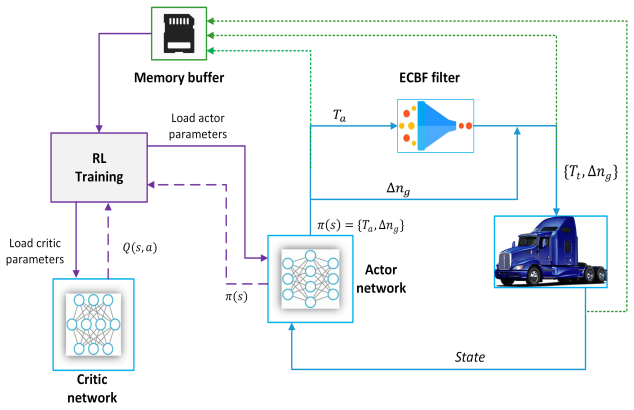


图 1. 用于自适应巡航控制的强化学习智能体训练

通过奖励函数衡量这些动作的适用性。奖励旨在完成范围内任务和范围外任务，并通过奖励权重（w）调节每项任务中的不同性能目标。当感知范围内不存在车辆（$z > z_{sr}$）时，如公式（11）所示，奖励结构要求车辆保持驾驶员设定速度，同时兼顾燃油消耗和扭矩变化平顺性。当感知范围内存在车辆时，奖励则旨在与前车保持较近距离，如公式（12）所示。在这种近距离情况下，除考虑扭矩变化平顺性和燃油消耗外，奖励 r_{os} 还会抑制本车超过驾驶员需求速度（v_{set}）的超速行为。通过加入由 w_g加权的挡位奖励项，可以减轻换挡振荡及其导致的车辆运行不平顺。

$$r = w_v 0.1^{\frac{|v_h - v_{set}|}{V_{rel,max}}} + w_f 0.1^{\frac{\dot{m}_f}{\dot{m}_{f,max}}} + w_T 0.1^{\frac{|\Delta T_e|}{T_{e,max}}} + w_g 0.1^{\frac{|\Delta n_g|}{n_{g,max}}} \quad (11)$$

$$r = w_z 0.1^{\frac{Z}{Z_{sr}}} + w_f 0.1^{\frac{\dot{m}_f}{\dot{m}_{f,max}}} + w_T 0.1^{\frac{|\Delta T_e|}{T_{e,max}}} + w_g 0.1^{\frac{|\Delta n_g|}{n_{g,max}}} + r_{os} \quad (12)$$

其中 $r_{os} = w_{os}\ if\ v_h \leq v_{set},\quad else: \quad r_{os} = w_{os} 0.1^{\frac{v_h - v_{set}}{v_{rel,max}}}$，$\dot{m}_f$ 为燃油消耗率，T_e 为发动机扭矩。为适应连续的牵引扭矩和离散的挡位选择，混合最大后验策略优化（HMPO）被认为是适合强化学习训练的算法（Kerbel 等人，2022；Neunert 等人，2020；Abdolmaleki 等人，2018）。与先进的近端策略优化（PPO）（Schulman 等人，2017）和信赖域策略优化（TRPO）（Schulman 等人，2015）算法一样，它具有可扩展性和鲁棒性；此外，由于它采用离策略方式，因此能够高效利用数据，适用于现实世界中的强化学习自适应巡航控制训练。强化学习智能体由演员网络（参数化为 $\boldsymbol{\theta}$）和评论家网络（参数化为 $\boldsymbol{\phi}$）组成，前者根据给定状态 $\pi(s|\boldsymbol{\theta})$ 确定控制策略，后者则通过提供相应的动作价值 $Q(s,\ a|\boldsymbol{\phi})$ 来评估这些动作。演员网络输出高斯分布的均值和方差，并从中采样得到牵引扭矩，

（13）。此外，输出层采用 softmax 激活函数，为换挡决策提供三个选项，分别对应可用的换挡操作 $\Delta n = \{1, 0, -1\}(upshift,\ nochange,\ downshift)$。随后使用分类采样来获得换挡策略（14）。假设连续策略 $\pi_{\boldsymbol{\theta}}^{T}(T_a|s)$ 与离散策略 $\pi_{\boldsymbol{\theta}}^{g}(\Delta n_g|s)$ 相互独立，则总策略可因式分解为（15），用于组合动作$a = \{T_a, \Delta n_g\}$。

$$\pi_{\boldsymbol{\theta}(T_a|s)}^{T} = \mathcal{N}\left(\mu_{\boldsymbol{\theta}}(s), \sigma_{\boldsymbol{\theta}}^{2}(s)\right) \quad (13)$$

$$\pi_{\boldsymbol{\theta}}^{g}(\Delta n_g|s) = Cat(\alpha_{\boldsymbol{\theta}}(s)), \forall s \sum_{k=1}^{3} \alpha_{k,\boldsymbol{\theta}}(s) = 1 \quad (14)$$

$$\pi_{\boldsymbol{\theta}}(a|s) = \pi_{\boldsymbol{\theta}}^{T}(T_a|s)\, \pi_{\boldsymbol{\theta}}^{g}(\Delta n_g|s)) \quad (15)$$

在策略改进阶段，MPO 从 Q 函数中针对不同动作进行采样，并更新演员网络参数，以输出使动作价值 $Q(s,\ a)$ 最大化的动作。为此，使用期望最大化算法优化采取最优动作的似然函数（参见 Neunert 等人（2020）；Abdolmaleki 等人（2018））。训练的策略评估阶段通过最小化当前 $Q_\theta(s,\ a,\ \boldsymbol{\phi})$ 与由 Retrace 采样 Q_t^{ret} (Munos 等人，2016) 定义的目标之间的平方损失，来拟合评论家网络的 Q 函数 $Q_\theta(s,\ a,\ \phi)$，其参数为 $\boldsymbol{\phi}$。

$$\min_{\boldsymbol{\phi}} L(\boldsymbol{\phi}) = \min_{\boldsymbol{\phi}} \mathbb{E}_{(s,a)\sim\mathcal{R}}\left[Q_{\boldsymbol{\theta}}(s,a|\boldsymbol{\phi}) - Q_t^{ret}\right]^2 \quad (16)$$

4. 结果与讨论

上述带有ECBF安全滤波器的强化学习自适应巡航控制应用于中型卡车模型，涵盖城市和高速公路驾驶工况。演员网络和评论家网络均由三个隐藏层构成，每层包含256个节点。仿真采用一辆10挡自动化手动变速器（AMT）卡车，其重量范围为5至10吨。城市驾驶训练期间，前车遵循联邦测试程序（FTP-75）驾驶循环；高速公路驾驶训练则依次采用高速公路燃油经济性测试（HWFET）和ArtMw130驾驶循环的组合（Barlow等，2009）。训练完成后，我们将使用不同的驾驶循环进行评估，具体如下文所述。

在每个仿真步骤中，如图1所示，演员网络会针对给定状态提出扭矩和挡位动作，再由ECBF安全层进行过滤。随后，车辆环境执行安全动作，并计算相应的奖励。为适应各项任务的不同目标，奖励函数针对范围内情况采用权重 $[w_v = 0.675, w_f$ =0.175, $w_T = 0.075, w_g = 0.075]$，针对范围外情况采用权重 $[w_z$ =0.325, $w_f = 0.175,$ $w_{os} = 0.35, w_T = 0.075, w_g = 0.075]$。状态、动作和奖励会存入记忆缓冲区，之后使用这些数据的批次通过HMPO算法训练网络。为了避免强化学习学会特定的驾驶循环，车辆会以随机间距初始化，并在前车的速度曲线上添加噪声。通过在训练回合内及回合之间改变卡车重量，考虑重量波动。

训练期间，由于按照第2节谨慎选择了增益向量 $K_\alpha = [0.2, 5]$ ，车辆从未发生碰撞，也从未进入安全距离 z_0以内。随着训练的进行，强化学习逐渐学会在超出范围时以接近驾驶员设定速度的状态行驶，并在范围内时越来越紧密地跟随前车。尽管没有提供发动机效率图，但从英里/加仑随训练提升可以看出，强化学习网络最终学会了燃油效率最优的挡位和扭矩动作。

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	演员、评论家的学习率	$10^{-4}, 10^{-5}$
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

尽管这种方法在安全关键系统中并不实用，但为了与基于 ECBF 的安全过滤进行比较，本文考虑采用奖励塑形来保障安全。当主车接近到小于最小安全距离限制 z_0 时，会在奖励函数中加入 $r_s = -1$ 的惩罚；若发生碰撞，惩罚则增至 $r_c = -10$。由于这些安全违规惩罚，随着训练进行，不安全动作会逐渐减少，最终智能体学会在确保安全的同时最大化奖励。除了奖励塑形方法外，还采用传统的 PID 自适应巡航控制作为基准；与强化学习的情况类似，该方法将控制分为范围内和范围外两个阶段进行设计（Canale 和 Malan，2003）。PID 控制器给出牵引扭矩 T_t 请求，并根据给定牵引扭矩和车辆速度下燃油消耗率最低的挡位选择最佳挡位（Yoon 等人，2020；Kerbel 等人，2022）。

训练完成基于 ECBF 的强化学习自适应巡航控制后，在城市和高速公路驾驶条件下，对其性能进行评估，并与 9 吨卡车上的 PID 自适应巡航控制和采用奖励塑形的强化学习自适应巡航控制进行比较。在城市工况下，前车遵循 ArtUrban 驾驶循环，驾驶员需求速度 v_{set} 设为 15 m/s。类似地，高速公路驾驶采用 25 m/s 的 v_{set}；为更好地捕捉高速公路工况下不同的速度曲线，前车遵循 ArtRoad 和 ARTMw150 的组合。在这两种工况下，车辆之间的初始间距均为 1500 m 。

在这两种驾驶条件下，强化学习自适应巡航控制均成功实现了范围内和范围外的目标，最重要的是，遵守了安全约束。图 2 显示，在模拟的大部分时间里，强化学习自适应巡航控制的速度曲线与 PID 自适应巡航控制相近。不过，在挡位选择方面，强化学习自适应巡航控制倾向于使用较高挡位。如表 2 所示，在高速公路驾驶中，强化学习自适应巡航控制的英里/加仑提升了 8.3%；而在城市驾驶中，其英里/加仑比 PID 自适应巡航控制基准高 7.9%。当前车处于范围内时，强化学习自适应巡航控制跟车更近，因此较不容易受到加塞的影响，表现为较低的平均范围内

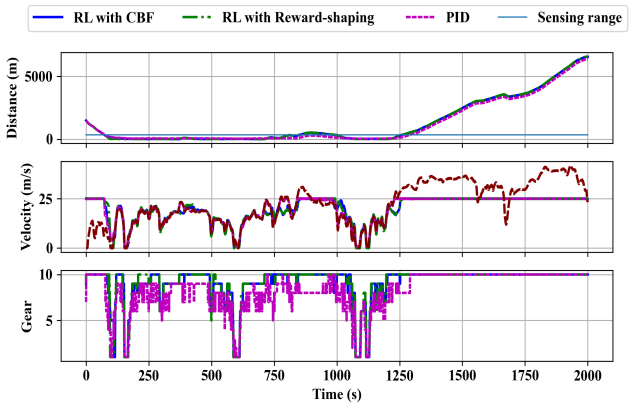


图 2. 高速公路驾驶中强化学习和 PID 自适应巡航控制器的间距、速度和挡位曲线仿真。

间距 z_{ir}。此外，可以看到，采用 ECBF 过滤器的强化学习自适应巡航控制与采用奖励塑形的强化学习自适应巡航控制，在燃油经济性和范围内跟车表现方面达到相当水平。

表 3 展示了重量变化情况下的性能比较，其中车辆重量范围为 5 至 10 吨。在给定的重量范围内，强化学习自适应巡航控制的英里/加仑始终高于 PID 自适应巡航控制，且间距未受到显著影响。

表 2. PID 自适应巡航控制、基于 ECBF 的强化学习自适应巡航控制与采用奖励塑形的强化学习自适应巡航控制的性能比较

	高速公路驾驶			城市驾驶		
ACC	PID	RL	RL	PID	RL	RL
安全层	-	ECBF	奖励塑形	-	ECBF	奖励塑形
MPG	8.6 (-)	9.3 (8.31%)	9.31 (8.37%)	6.8 (-)	7.35 (7.9%)	7.38 (8.4%)
$Z_{ir}(m)$	95	74	73	42	39	38

5. 结论

本文采用基于指数控制障碍函数的安全过滤器，通过将学习探索投影到安全集，为基于强化学习的自适应巡航控制系统注入安全性。由于实际系统的动作范围有限，我们提出了一种方法，通过考虑最坏情况进行前向仿真，以验证给定指数控制障碍函数设计的安全性。经过该指数控制障碍函数过滤后，强化学习自适应巡航控制提出的牵引扭矩和换挡动作将在模拟车辆环境中执行，并记录相应的奖励。强化学习网络采用混合最大 A

表 3. 车辆质量波动下 PID 自适应巡航控制和强化学习自适应巡航控制的随机化性能

	重量（吨）	5	6	7	8	9	10
RL with ECBF	MPG	10.58 (10.9%)	10.38 (11.6%)	9.99 (9.6%)	9.61 (8.3%)	9.3 (8.31%)	8.95 (7.6%)
	$Z_{ir}(m)$	67	69	73	75	74	77
PID	MPG	9.54	9.3	9.11	8.87	8.6	8.32
	$Z_{ir}(m)$	95	95	94	95	95	96

后验策略优化（HMPO）算法进行训练，该算法兼容连续的牵引扭矩动作和离散的换挡动作。对一辆中型卡车的评估表明，强化学习自适应巡航控制实现了速度目标，且最重要的是遵守了安全约束。与 PID 自适应巡航控制相比，当高速公路驾驶工况下前车依次跟随 ArtRoad 和 ARTMw150 驾驶循环时，强化学习自适应巡航控制的英里/加仑提高了 8.3%；当城市驾驶工况下前车跟随 ArtUrban 驾驶循环时，则提高了 7.9%。此外，强化学习自适应巡航控制能够学会应对重量波动，并在车辆整个重量范围内保持高性能。

当前算法的训练和评估是在标准驾驶循环上进行的。未来的工作将侧重于使用随机交通数据和测量噪声，在更加真实的驾驶条件下评估强化学习自适应巡航控制的性能和鲁棒性。此外，未来的工作还将研究在指数控制障碍函数设计中考虑不确定性时不那么保守的方法（而非最坏情况方法）。

参考文献

Abdolmaleki, A., Springenberg, J.T., Tassa, Y., Munos, R., Heess, N., and Riedmiller, M.（2018）。最大后验策略优化。第*6*届国际学习表征会议。Altman, E.（1999）。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）。控制屏障函数：理论与应用。*2019*年第*18*届欧洲控制会议，*ECC 2019*，3420–3431。Ames, A.D., Grizzle, J.W., and Tabuada, P.（2014）。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-Febru（February），6271–6278。Barlow, T.J., Latham, S., Mccrae, I.S., and Boulter, P.G.（2009）。用于测量道路车辆排放的驾驶循环参考手册。Canale, M. and Malan, S.（2003）。基于PID的自适应巡航控制启停系统的鲁棒设计。*IFAC*会议论文集，36(18)，333–338。Chamraz, S. and Balogh, R.（2018）。自适应巡航控制（ACC）的两种设计方法。第*29*届国际控制论与信息学会议，*K and I 2018*，2018-Janua（2），1–6。Cheng, R., Orosz, G., Murray, R.M., and Burdick, J. W.（2019）。通过障碍函数为安全关键型连续控制任务实现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., Padu-raru, C., and Tassa, Y.（2018）。连续动作空间中的安全探索。Hsu, S.C., Xu, X., and Ames, A.D.（2015）。基于控制屏障函数的二次规划及其在双足机器人行走中的应用。美国控制会议论文集，2015-July，4542–4548。Kerbel, L., Ayalew, B., Ivanco, A., and Loiselle, K.（2022）。基于深度强化学习的驾驶辅助节能驾驶与变速器控制。Lang, D., Stanger, T., Schmied, R., and del Re, L.（2014）。预测协同自适应巡航控制：燃油消耗效益与可实施性。163–178。Li, G. and G¨orges, D.（2020）。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。Li, Z.（2021）。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*。Marsden, G., McDonald, M., and Brackstone, M.（2001）。增进对自适应巡航控制的理解。交通研究*C*辑：新兴技术，9(1)，33–51。Munos, R., Stepleton, T., Harutyunyan, A., and Bellemare, M.G.（2016）。安全且高效的离策略强化学习。神经信息处理系统进展，1054–1062。doi:10.48550/arxiv.1606.02647。URL https://arxiv.org/abs/1606.02647v2。Naus, G., Van Den Bleek, R., Ploeg, J., Scheepers, B., Van De Molengraft, R., and Steinbuch, M.（2008）。自适应巡航控制启停的显式模型预测控制设计与性能评估。美国控制会议论文集，224–229。Neunert, M., Abdol maleki, A., Wulfmeier, M., Lampe, T., Springenberg, J.T., Hafner, R., Romano, F., Buchli, J., Heess, N., and Riedmiller, M.（2020）。机器人学中混合控制的连续—离散强化学习。（CoRL）。Nguyen, Q. and Sreenath, K.（2016）。用于强制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-July（3），322–328。Radke, F. and Isermann, R.（1987）。一种采用逐步参数优化的参数自适应PID控制器。*Automatica*，23(4)，449–457。Schulman, J., Levine, S., Moritz, P., Jordan, M., and Abbeel, P.（2015）。信赖域策略优化。第*32*届国际机器学习会议，*ICML 2015*，3，1889–1897。Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O.（2017）。近端策略优化算法。1–12。Srinivasan, K., Eysenbach, B., Ha, S., Tan, J., and Finn, C.（2020）。学习如何确保安全：采用安全评论器的深度强化学习。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）。恢复强化学习：借助学习得到的恢复区域实现安全强化学习。*IEEE*机器人与自动化快报，6(3)。Xu, X., Grizzle, J.W., Tabuada, P., and Ames, A.D.（2018）。车道保持与自适应巡航控制组合的正确性保证。*IEEE*自动化科学与工程汇刊，15(3)，1216–1229。Yang, Z., Wang, Z., and Yan, M.（2021）。基于模型预测控制和自抗扰控制的自适应巡航控制系统优化设计。*Actuators 2021*，第*10*卷，第*110*页，10(6)，110。Yoon, D.D., Ayalew, B., Ivanco, A., and Loiselle, K.（2020）。用于……的预测性动能管理

重型车辆的附加驾驶辅助生态驾驶。*IET* 智能交通系统，
14(13)，1824–1834。