Preprints, 10th IFAC International Symposium on
Advances in Automotive Control
Columbus, Ohio, USA, August 28-30, 2022

Safety Filtering for Reinforcement Learning-based Adaptive Cruise Control

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

* *Automotive Engineering, Clemson University, Greenville, SC 29607, USA (hhailem, beshah, lsutto2)@clemson.edu.*
** *Allison Transmission Inc., One Allison Way, Indianapolis, IN, 46222, USA (andrej.ivanco, keith.loiselle)@allisontransmission.com*

Abstract: Reinforcement learning (RL)-based adaptive cruise control systems (ACC) that learn and adapt to road, traffic and vehicle conditions are attractive for enhancing vehicle energy efficiency and traffic flow. However, the application of RL in safety critical systems such as ACC requires strong safety guarantees which are difficult to achieve with learning agents that have a fundamental need to explore. In this paper, we derive control barrier functions as safety filters that allow an RL-based ACC controller to explore freely within a collision safe set. Specifically, we derive control barrier functions for high relative degree nonlinear systems to take into account inertia effects relevant for commercial vehicles. We also outline an algorithm for accommodating actuation saturation with these barrier functions. While any RL algorithm can be used as the performance ACC controller together with these filters, we implement the Maximum A Posteriori Policy Optimization (MPO) algorithm with a hybrid action space that learns fuel optimal gear selection and torque control policies. The safety filtering RL approach is contrasted with a reward shaping RL approach that only learns to avoid collisions after sufficient training. Evaluations on different drive cycles demonstrate significant improvements in fuel economy with the proposed approach compared to baseline ACC algorithms.

Keywords: Adaptive cruise control, Safe reinforcement learning, Safety filtering, Control barrier functions

1. INTRODUCTION

Adaptive cruise control (ACC) systems are one of the increasingly prevalent driver assistance systems for modern vehicles. An ACC system uses radar, computer vision, or laser to understand the vehicle's surrounding and make control decisions. When another vehicle or object is not in the sensing range, ACC compensates for road grade, friction, and aerodynamic resistances to maintain a speed set by the driver. When another car or object is in front, the ACC makes decisions to prevent collision and follow the preceding vehicle as close as possible to avoid cut-ins. ACC has been shown to decrease a drivers' work load, make traffic flows efficient and driving safer (Marsden et al., 2001; Lang et al., 2014).

An effective ACC system should balance the traffic condition of the road, the vehicle performance, and the driver's demanded velocity. Currently available PID-based ACC systems (Canale and Malan, 2003; Chamraz and Balogh, 2018) and proposed MPC-based approaches (Naus et al., 2008; Yang et al., 2021) are often tuned to balance this trade-off for various operating environments. Although 'adaptive' or gain-scheduled versions (Radke and Isermann, 1987) can be sought, the fixed structure of these approaches limits full adaptation throughout the lifetime of the vehicle. Furthermore, MPC-based ACC also has to find a reliable way of predicting the motion of the leading vehicle for a future horizon. On the other hand, data-driven reinforcement learning (RL) approaches offer a mechanism to continuously customize to traffic, road and vehicle conditions without a predefined control architecture (Li and Görges, 2020). In this work, we consider applications of RL-based ACC to commercial vehicles. In addition, while traditional ACC is primarily about the two tasks of speed tracking and maintaining a safe gap, we consider RL-based ACC (RL ACC for short) to explicitly optimize fuel economy via gear selection and torque control policies.

Despite the potential benefits of adaptability and improved performance, RL ACC faces critical safety challenges. These derive from the needs of RL algorithms to explore in order to learn the optimal policies. RL learns how good the given state-action pair is after experiencing it, but for applications like vehicle control, exploration in an unsafe domain is unacceptable even during (on-road) training of the RL algorithms. However, thanks to recent progress in safe RL, different approaches are suggested to encourage or limit the exploration only in the safe domain. We briefly mention a few of them. Reward shaping approaches put large penalties into the performance objective function if constraints are violated. On the other hand, constrained Markov decision process (CMDP) approaches assign safety constraint costs to each state-action pair and limit the total safety constraint cost of a trajectory to be lower than a certain threshold (Altman, 1999). The

reward shaping and CMDP approaches are implemented on the performance controller itself to encourage respecting safety constraints but they do not guarantee safety. Another set of approaches involve the use of safety filters that impose hard constraints. Such approaches separate the performance-oriented RL controller, whose only aim is to optimize the system's performance objective function, from the safety filters, which project the unsafe actions proposed by the performance controller into the safe set. The safety filters determine the safety condition of the given state-action pair using the dynamical model of the system, or they use offline data to learn constraints (Dalal et al., 2018) and safety indexes (Thananjeyan et al., 2021; Srinivasan et al., 2020). In this paper, we pursue dynamical model-based safety guarantees to construct the safe set in such a way that gives the RL performance controller the freedom to explore within the safe boundaries. As its training progresses, the RL performance controller eventually learns the safety boundaries and ceases to demand unsafe actions (Thananjeyan et al., 2021). Note that even though it does not interfere in the inner workings, the safety filter affects control performance by dictating where the performance controller could operate.

Of the model-based approaches to designing safety filters, control barrier functions (CBFs) offer light computation and scalability (Li, 2021). A CBF guarantees safety by making the controller work in the invariant safe-set defined by a superlevel set of a continuously differentiable function $h(x) : R^n \to R$. The action selected by the performance controllers are projected into the safe set in such a manner that the proposed actions are modified minimally (Ames et al., 2019), and no unsafe actions are passed to the controlled system. Different approaches could be pursued to specify CBFs with their pros and cons. The intuitive one is to come up with a handcrafted CBF considering the dynamics of the system and the action bounds associated with it (Xu et al., 2018; Ames et al., 2014; Cheng et al., 2019). In collision avoidance problems, for instance, the CBF can be derived by considering the maximum deceleration that the system could exert to close a distance gap. When possible, it is also desirable to progressively widen the safe set to get the maximal safe domain, a task currently possible with polynomial plant dynamics and polynomial CBFs via sum-of-squares (SOS) programming (Chamraz and Balogh, 2018). Another approach that is tailored to high relative degree nonlinear dynamical systems such as those involving inertia effects is the use of exponential CBF (ECBF) (Nguyen and Sreenath, 2016). In this work, we derive ECBFs to work as safety filters with our RL-ACC controllers, thereby taking explicit considerations of inertia effects that are important for commercial vehicles that experience large changes in loading.

The main contributions of this paper are then the derivation and demonstration of CBF-based safe RL-ACC approach for commercial vehicles that optimizes fuel economy. While we derive ECBFs for safety certification, we note that straight ECBFs (or CBFs in general) assume unbounded actions, and in their natural form, they might request actions that are not feasible for the vehicle's powertrain to meet. We therefore put forward a method to provide a safety guarantee for a given parameters of ECBF within the vehicle action limits. Our performance RL-ACC coordinates traction torque control and gear decisions considering fuel consumption optimization objectives. The RL ACC augmented with the safety certificate is trained and evaluated on different driving cycles, and the vehicle performance is compared with an RL ACC with reward-shaping approach to safe RL, as well as with a conventional PID-based ACC.

The rest of the paper is organized as follows. Section 2 describes our derivation of the ECBF as safety filters for ACC and detail how we address actuation constraints within them. Section 3 describes the algorithmic details of our performance RL-ACC. Section 4 discusses results and discussions, and Section 5 concludes the paper.

2. SAFETY FILTER FOR ACC

We briefly review the definition of CBFs as follows. Details are given in Hsu et al. (2015). Consider a nonlinear control affine system:

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

where f and g are locally Lipschitz, $x \in \mathcal{R}^n$ is the system state, $u \in \mathcal{R}^m$ is the control inputs. Assume a safe set defined by $\mathcal{C} = \{x \in \mathcal{R}^n | h(x) \geq 0\}$, where $h : \mathcal{R}^n \to \mathcal{R}$ is a continuously differentiable function. Then h is a control barrier function (CBF) if there exists an extended class κ_∞ function α such that for all $x \in Int(\mathcal{C}) = \{x \in \mathcal{R}^n : h(x) > 0\}$:

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

For high relative degree nonlinear affine systems, feedback linearization could be used to develop exponential CBFs (ECBF) as detailed in Nguyen and Sreenath (2016). This is accomplished by transforming (input-output linearizing) the high relative degree nonlinear systems into a virtual linear systems with new state variable $\eta_b := [h(x), \dot{h}(x), \cdot \cdot \cdot, h^r(x)]^T$, input μ and output $h(x)$:

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

where F and G are matrices representing an integrator chain, and $C = [1,\ 0,\ \cdot \cdot \cdot, 0]$. A state feedback controller can be designed for the transformed system as: $\mu = -K_\alpha \eta_b$ with a suitable gain vector K_α that makes $F - GK_\alpha$ Hurwitz. For a system with relative degree r, μ is also r^{th} derivative of the output $h(x)$, $\mu = L_f^r h(x) + L_g L_f^{r-1} h(x) u$. If there exists a state feedback gain K_α that makes $\mu \geq -K_\alpha \eta_b(x)$ for all states, then one can show that $h(x)$ is an exponential control barrier function (see Nguyen and Sreenath (2016)).

The ACC part of the present problem is modeled with the state variables of separation distance z, the velocity of the host vehicle v_h and velocity of the leading vehicle v_l. The corresponding state equations are:

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

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

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

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

where F_r is the total resistance force including gravitational, rolling and aerodynamic resistances, and T_t is the traction torque at the wheels.The parameters c_d, f, θ, m_v , ρ, A_v, r_w, a_l are aerodynamic coefficient, rolling resistance coefficient, road grade, mass of the vehicle, density of air, frontal area of the vehicle, radius of the wheels, and acceleration of the leading vehicle, respectively.

We observe that the above model can be readily put in the control affine form (1). Given a collision safety objective, we seek the separation distance z to always be above a specified minimum inter-vehicle distance z_0. To this end, we define the control barrier function (CBF) as the output $h(x) = z - z_0$. Considering that the control actuation is the traction torque T_t, we have a control affine system of relative degree two. In physical terms, the safety objective is on position while traction torque directly manipulates acceleration. Inertia effects come into play and must be accounted for. The input-output linearization into the form (3) then gives:

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

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

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

We now compute some bounds for the given control input μ considering actuation limits on the traction torque (T_{min} and T_{max}). For a given acceleration of the preceding vehicle (a_l) and velocity of the host (v_h), the feasible bounds of μ are given as

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

For a given gain vector $K_\alpha = [k_{\alpha 1},\ k_{\alpha 2}]$, ECBF guarantees safety if the proposed state feedback control, $-k_{\alpha 1}(z - z_0) - k_{\alpha 2}(v_l - v_h)$, is within the virtual linear system action bound $[\mu_{Tmax},\ \mu_{Tmin}]$. In general application cases, however, this bound may not be respected. Nevertheless, if K_α is chosen so that the poles are placed sufficiently to the left in s-plane, the above ECBF could still bound the safe set. Safety assurance for such pole selections could be achieved by investigating the evolution of the CBF control term $h(x)$ in worst-case situation where the linear virtual model is initialized with extreme possible $\eta_{0,xrm}$, and then the possible limiting torque actions are applied. For a given minimum separation distance target and maximum downhill road grade, this is equivalent to applying the maximum possible traction torque output of the performance RL-ACC agent, with the host vehicle model (of largest loading) initialized in with the maximum possible velocity while the preceding vehicle is under its maximum deceleration. This extreme conditions gives the feasible μ bounds as $\mu_{Tmin-xrm}$ and $\mu_{Tmax-xrm}$ using equations (9).

To capture the evolution of $h(x)$ under these extreme conditions, a simulation rollout is discretized into timestep Δt, and the action μ (saturated with $\mu_{Tmin-xrm}$ and $\mu_{Tmax-xrm}$) held piecewise constant. Algorithm 1 shows how this implemented by integrating the virtual system (3). If the $h(x)$ from this simulation is positive at infinity (or after some finite time), the selected K_α guarantees safety. Otherwise, the K_α need to be changed until this is satisfied.

Algorithm 1 An algorithm to enforce system bounds on a virtual linear system

$\eta \leftarrow \eta_0$
$\mu \leftarrow \mu_0$
while $t \le t_\infty$ **do**
  $t \leftarrow t + \Delta t$
  if $\mu < \mu_{Tmax-xrm}$ **then**
    $\mu \leftarrow \mu_{Tmax-xrm}$
  else if $\mu > \mu_{Tmin-xrm}$ **then**
    $\mu \leftarrow \mu_{Tmin-xrm}$
  end if
  $h(x(t)) \leftarrow C(e^{F\Delta t}\eta_0 + e^{F\Delta t}\int_0^{\Delta t} e^{-F\tau} G\mu d(\tau))$
  $\mu \leftarrow -k_{\alpha 1} h(x) - k_{\alpha 2}\dot{h}(x)$
  $\eta_0 \leftarrow \begin{bmatrix} h(x) \\ \dot{h}(x) \end{bmatrix}$
end while

Once the suitable gain vector K_α are selected, the ECBF safety constraint enforces safety by projecting the action proposed by the outputs of the RL controller's actor network $T_a(s)$ (see next section) to the control traction torque T_t in a way that introduces minimal changes to it. This is done by posing and solving the quadratic program:

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

3. VEHICLE ENVIRONMENT AND RL ACC

The powertrain controller is modeled as Markov decision process (MDP) consisting of states s, actions a, a reward function $r(s, a)$, and discounting factor γ. The probability of action choices is policy $\pi(a|s, \boldsymbol{\theta})$ where $\boldsymbol{\theta}$ denotes the parameters of the deep neural network used to approximate the policy. The host vehicle velocity v_l, the relative velocity between the preceding and host vehicles v_{rel}, the separation distance between the vehicles z, the gear n_g, the mass of the vehicle m_v, the road grade θ, the driver demanded velocity v_{set} and a flag to show if the vehicle is in ACC sensor range f constitute the states of the RL agent, $s = \{v_l, v_{rel}, z, n_g, m_v, \theta, v_{set}, f\}$. The RL performance controller is designed to perform both traction torque T_a control and gear change selection Δn_g, i.e. $a = \{T_a, \Delta n_g\}$. As shown in Fig.1, the proposed T_a is filtered by the ECBF safety layer to safe traction torque demand T_t (10). The engine torque and engine speed that brings about this wheel traction torque are then calculated utilizing transmission ratios of the selected gear and the final drive, and the associated fuel rate is read from the fuel map. Notice that while the RL controllers actions are T_a and Δn_g, the ECBF safety filter does not use Δn_g in the safety constraint. However, taking into account that gear selection is crucial for fuel economy and driver accommodation, it is an integral part of the RL performance controller.

The filtered traction torque T_t and the gear change Δn_g actions are implemented in the vehicle environment, and the suitability of the actions is measured by the reward function. The reward is designed to accomplish the in range and out of range tasks, and different performance

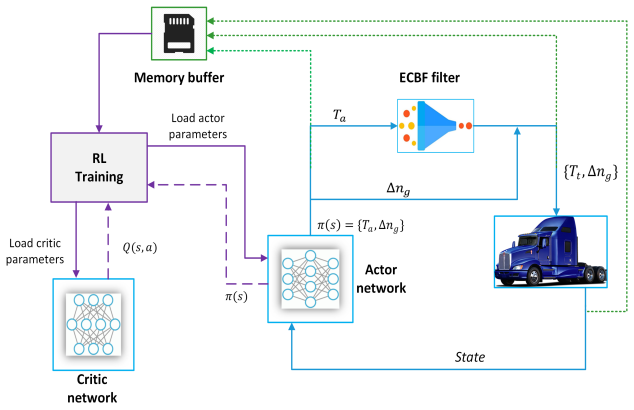

Fig. 1. Training RL agent for ACC

objectives within each task are tuned by reward weights (w). When there is not a vehicle present in the sensing range ($z > z_{sr}$), as shown in (11), the reward structure requires the vehicle to maintain the driver-set velocity and concurrently balances the fuel consumption and smooth torque change considerations. When there is a vehicle in the sensing range, on the other hand, the reward aims to maintain a close distance with the preceding vehicle, as shown in (12). In such proximity, in addition to smooth torque change and fuel consumption considerations, the reward r_{os} discourages the host vehicle from overspeeding beyond the driver demanded velocity (v_{set}). Gear hunting and the associated rough vehicle operation are mitigated by including a gear reward term weighted by w_g.

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

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

where $r_{os} = w_{os}\ if\ v_h \leq v_{set},\ else: r_{os} = w_{os} 0.1^{\frac{v_h - v_{set}}{v_{rel,max}}}$, $\dot{m}_f$ is the fuel rate and T_e is the engine torque.

To accommodate the continuous traction torque and the discrete gear selection, Hybrid Maximum A Posteriori Policy Optimization (HMPO) is found to be a good fit for the RL training algorithm (Kerbel et al., 2022; Neunert et al., 2020; Abdolmaleki et al., 2018). In addition to being scalable and robust like state of the art Proximal Policy Optimization (PPO) (Schulman et al., 2017) and Trust-Region Policy Optimization (TRPO) (Schulman et al., 2015) algorithms, the fact that it is off-policy makes it data efficient to apply it to the real world RL ACC trainings. The RL agent comprises of an actor (parameterized by $\boldsymbol{\theta}$) and a critic (parameterized by $\boldsymbol{\phi}$) networks, in which the former determines the control policy for a given state $\pi(s|\boldsymbol{\theta})$ and the latter evaluates these actions by providing the associated action values $Q(s, a|\boldsymbol{\phi})$. The actor network outputs mean and variance of a gaussian distribution, from which traction torque is sampled (13). In addition to that, it uses softmax activation at the output layer with three choices for the gear change decision, analogous to the available gear changes $n = \{1, 0, -1\}(upshift, nochange, downshift)$. Categorical sampling is then used to obtain the gear change policy (14). Assuming independence between the continuous $\pi^T_{\boldsymbol{\theta}}(T_a|s)$ and discrete $\pi^g_{\boldsymbol{\theta}}(\Delta n_g|s)$ policies, the total policy could be factorized as (15) for combine action $a = \{T_a, \Delta n_g\}$.

$$\pi^T_{\boldsymbol{\theta}(T_a|s)} = \mathcal{N}\left(\mu_{\boldsymbol{\theta}}(s), \sigma^2_{\boldsymbol{\theta}}(s)\right) \tag{13}$$

$$\pi^g_{\boldsymbol{\theta}}(\Delta n_g|s) = Cat(\alpha_{\boldsymbol{\theta}}(s)), \forall s \sum_{k=1}^{3} \alpha_{k,\boldsymbol{\theta}}(s) = 1 \tag{14}$$

$$\pi_{\boldsymbol{\theta}}(a|s) = \pi^T_{\boldsymbol{\theta}}(T_a|s)\, \pi^g_{\boldsymbol{\theta}}(\Delta n_g|s)) \tag{15}$$

In the policy improvement phase, MPO samples from the Q-function for different actions and update the actor-network parameters to output actions that maximize the action values $Q(s, a)$. This is accomplished by optimizing the likelihood function of acting optimally using the expectation-maximization algorithm (see Neunert et al. (2020); Abdolmaleki et al. (2018)). The policy evaluation phase of the training fits the Q-function $Q_\theta(s, a, \phi)$ of the critic network, with parameters $\boldsymbol{\phi}$, by minimizing the square loss of the current $Q_\theta(s, a, \boldsymbol{\phi})$ and a target defined by retrace sampling Q^{ret}_t (Munos et al., 2016).

$$\min_{\boldsymbol{\phi}} L(\boldsymbol{\phi}) = \min_{\boldsymbol{\phi}} E_{(s,a)\sim\mathcal{R}}\left[Q_{\boldsymbol{\theta}}(s, a|\boldsymbol{\phi}) - Q^{ret}_t\right]^2 \tag{16}$$

4. RESULTS AND DISCUSSIONS

The above RL ACC with the ECBF safety filter is applied to a model of medium duty truck in urban and highway driving conditions. The actor and critic networks are constructed with three hidden layers, and each layer consists of 256 nodes. The simulation uses a 10-speed automated manual transmission (AMT) truck that has a 5 to 10 tons weight range. The preceding vehicle follows Federal Test Procedure (FTP-75) drive cycle for the urban driving training, while for highway driving, a combination of Highway Fuel Economy (HWFET) and ArtMw130 cycles are used in succession (Barlow et al., 2009) . Once trained, we will use different drive cycles for evaluation as will be describe below.

In each simulation step, as shown in Fig.1, the actor network proposes the torque and the gear actions for a given state which will be filtered by the ECBF safety layer. The vehicle environment then executes the safe actions, and the associated rewards are calculated. To accommodate the different objectives of each task, the reward is structured with weights of $[w_v = 0.675, w_f = 0.175, w_T = 0.075, w_g = 0.075]$ for in range, and $[w_z = 0.325, w_f = 0.175, w_{os} = 0.35, w_T = 0.075, w_g = 0.075]$ for out of range conditions. The state, action and rewards are stored in the memory buffer, and afterward, batches of these data are used to train the networks using the HMPO algorithm. In order to prevent RL from learning the specific drive cycles, the vehicles are initialized in random separation distance along with the addition of noise to the velocity profile of the preceding vehicle. The weight fluctuations are considered by varying the truck weight within and between training episodes.

During training, because of the careful choice of the gain vector $K_\alpha = [0.2, 5]$ as per section 2, the vehicle never

crashes nor comes within safe distance z_0. As the training progresses, the RL learns to operate near the driver set velocity when it is out of range and follows the preceding vehicle more and more closely when it is in range. Even though it is not provided with the engine efficiency map, as exhibited by the improvement of MPG with training, the RL network eventually learns the fuel optimal gear and torque actions.

Table 1. Vehicle environment and RL hyperparameter setting

Vehicle Parameters		MPO Hyperparameters	
Mass	5 - 10 tons	Actor, critic learning rate	$10^{-4}, 10^{-5}$
A_u	$7.71m^2$	Dual constraint	0.1
C_d	0.08	Retrace steps	15
r_w	0.498	KL constraints $\epsilon_\mu, \epsilon_\sigma, \epsilon_d$	$0.1, 0.001, 0.1$
f	0.015	α_d, α_c	10
z_{sr}	350	γ	0.99

Even if it is not practical for safety critical systems, a reward shaping approach of safeguarding safety is considered to compare against the ECBF-based safety filtering. A penalty of $r_s = -1$ is added to the reward function when the host approaches closer than the minimum safe distance limit z_0 and, in the situation of a crash, the penalty is enlarged to $r_c = -10$. Due to these safety violation penalties, unsafe actions reduce with training, and eventually, the agent learns to maximize the reward safely. In addition to the reward shaping approach, the conventional PID ACC is used as a baseline which, like in the case of RL, is designed by dividing the control into phases for the in rage and out of range conditions (Canale and Malan, 2003). The traction torque T_t request is given by PID controller and an optimal gear is chosen based on the gear with the lowest fuel rate given the desired traction torque and vehicle velocity (Yoon et al., 2020; Kerbel et al., 2022).

After the RL ACC with ECBF is trained, its performance is evaluated and compared with PID ACC and RL ACC with reward shaping counterparts on a 9-ton truck in the urban and highway driving conditions. For the urban case, the preceding vehicle follows the ArtUrban drive cycle, and the driver demanded velocity v_{set} is set to be 15 m/s. Similarly, a v_{set} of 25 m/s is used for highway driving, and to better capture different velocity profiles in the highway situation, the preceding vehicle follows a combination of ArtRoad and ARTMw150. The initial separation distance between the vehicles is 1500 m in both cases.

In both driving conditions, the RL ACC successfully met the in range as well as out of range objectives and, most importantly, safety constraints are respected. Fig.2 shows the RL ACC has a similar velocity profile to its PID ACC counterpart for the most part of the simulation. However, when it comes to gear selection, the RL ACC tends to operate at higher gears. As summarized in Table 2, for the highway driving, the RL ACC exhibited an MPG improvement of 8.3%, whereas, in the case of the urban driving, it has 7.9% higher MPG than the PID ACC baseline. When the preceding vehicle is in range, the RL ACC is less susceptible to cut-in as it follows the preceding vehicle closer, shown by the lower mean in range separation distance z_{ir}. Moreover, it is possible to see that the RL ACC with ECBF filter and the RL ACC with

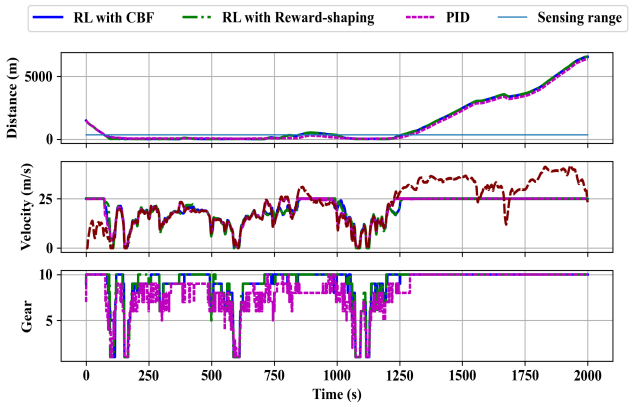

Fig. 2. Simulation of separation distance, velocity, and gear profiles of RL and PID ACC controllers in a highway driving.

reward shaping arrangements achieve equivalent levels of fuel economy and in range car following performances.

Table 3 shows the performance comparison with weight fluctuation in which the vehicle's weight ranges from 5 to 10-tons. The RL ACC's maintains higher MPG than the PID ACC throughout the given weight range, and the separation distance is not significantly influenced.

Table 2. Performance comparison between PID ACC, RL ACC with ECBF and RL ACC with reward shaping

	Highway driving			Urban driving		
ACC	PID	RL	RL	PID	RL	RL
Safety layer	-	ECBF	Reward shaping	-	ECBF	Reward shaping
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. CONCLUSION

In this paper, an exponential control barrier function-based safety filter is employed to instill safety into RL based ACC system by projecting the learning exploration to a safe set. Since practical systems operate with bounded actions, we proposed an approach to verify the safety of a given ECBF design by forward simulating in consideration of worst case scenarios. After being filtered by this ECBF, the traction torque and gear change actions proposed by the RL-based ACC are implemented on a simulated vehicle environment and the associated rewards are observed. The RL networks are trained using Hybrid Maximum A Posteriori Policy Optimization (HMPO) algorithm that accommodates the continuous traction torque and discrete

Table 3. Performance of PID ACC and RL ACC with vehicle mass fluctuation

	Weight (tons)	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

gear change actions. Evaluation on a medium-duty truck shows that the RL ACC fulfilled the velocity objectives and, most importantly, respected the safety constraints. As compared to PID ACC, the RL ACC augments MPG by 8.3% in highway driving conditions when the preceding vehicle follows a combination of ArtRoad and ARTMw150 drive cycles, and by 7.9% in urban driving conditions when the preceding vehicle follows ArtUrban drive cycle. Moreover, the RL ACC learns to handle weight fluctuations and maintains high performance throughout the vehicle's weight range.

The current algorithm training and evaluations are performed on standard driving cycles. Future work will focus on using randomized traffic data and measurement noise to assess the performance and robustness of RL ACC in even more realistic driving conditions. In addition, future work will also look at less conservative methods of accounting for uncertainties (not worst-case) in ECBF design.

REFERENCES

Abdolmaleki, A., Springenberg, J.T., Tassa, Y., Munos, R., Heess, N., and Riedmiller, M. (2018). Maximum a posteriori policy optimisation. *6th International Conference on Learning Representations*.

Altman, E. (1999). Constrained Markov Decision Processes .

Ames, A.D., Coogan, S., Egerstedt, M., Notomista, G., Sreenath, K., and Tabuada, P. (2019). Control barrier functions: Theory and applications. *2019 18th European Control Conference, ECC 2019*, 3420–3431.

Ames, A.D., Grizzle, J.W., and Tabuada, P. (2014). Control barrier function based quadratic programs with application to adaptive cruise control. *Proceedings of the IEEE Conference on Decision and Control*, 2015-Febru(February), 6271–6278.

Barlow, T.J., Latham, S., Mccrae, I.S., and Boulter, P.G. (2009). A reference book of driving cycles for use in the measurement of road vehicle emissions.

Canale, M. and Malan, S. (2003). Robust design of PID based ACC S and G systems. *IFAC Proceedings Volumes*, 36(18), 333–338.

Chamraz, S. and Balogh, R. (2018). Two approaches to the adaptive cruise control (ACC) design. *Proceedings of the 29th International Conference on Cybernetics and Informatics, K and I 2018*, 2018-Janua(2), 1–6.

Cheng, R., Orosz, G., Murray, R.M., and Burdick, J.W. (2019). End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019*, 3387–3395.

Dalal, G., Dvijotham, K., Vecerik, M., Hester, T., Paduraru, C., and Tassa, Y. (2018). Safe Exploration in Continuous Action Spaces.

Hsu, S.C., Xu, X., and Ames, A.D. (2015). Control barrier function based quadratic programs with application to bipedal robotic walking. *Proceedings of the American Control Conference*, 2015-July, 4542–4548.

Kerbel, L., Ayalew, B., Ivanco, A., and Loiselle, K. (2022). Driver Assistance Eco-driving and Transmission Control with Deep Reinforcement Learning.

Lang, D., Stanger, T., Schmied, R., and del Re, L. (2014). Predictive Cooperative Adaptive Cruise Control: Fuel Consumption Benefits and Implementability. 163–178.

Li, G. and Görges, D. (2020). Ecological Adaptive Cruise Control for Vehicles with Step-Gear Transmission Based on Reinforcement Learning. *IEEE Transactions on Intelligent Transportation Systems*, 21(11), 4895–4905.

Li, Z. (2021). Comparison between safety methods control barrier function vs. reachability analysis. *arXiv preprint arXiv:2106.13176*.

Marsden, G., McDonald, M., and Brackstone, M. (2001). Towards an understanding of adaptive cruise control. *Transportation Research Part C: Emerging Technologies*, 9(1), 33–51.

Munos, R., Stepleton, T., Harutyunyan, A., and Bellemare, M.G. (2016). Safe and Efficient Off-Policy Reinforcement Learning. *Advances in Neural Information Processing Systems*, 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). Explicit MPC design and performance evaluation of an ACC stop--go. *Proceedings of the American Control Conference*, 224–229.

Neunert, M., Abdolmaleki, A., Wulfmeier, M., Lampe, T., Springenberg, J.T., Hafner, R., Romano, F., Buchli, J., Heess, N., and Riedmiller, M. (2020). Continuous-Discrete Reinforcement Learning for Hybrid Control in Robotics. (CoRL).

Nguyen, Q. and Sreenath, K. (2016). Exponential Control Barrier Functions for enforcing high relative-degree safety-critical constraints. *Proceedings of the American Control Conference*, 2016-July(3), 322–328.

Radke, F. and Isermann, R. (1987). A parameter-adaptive PID-controller with stepwise parameter optimization. *Automatica*, 23(4), 449–457.

Schulman, J., Levine, S., Moritz, P., Jordan, M., and Abbeel, P. (2015). Trust region policy optimization. *32nd International Conference on Machine Learning, ICML 2015*, 3, 1889–1897.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal Policy Optimization Algorithms. 1–12.

Srinivasan, K., Eysenbach, B., Ha, S., Tan, J., and Finn, C. (2020). Learning to be Safe: Deep RL with a Safety Critic. 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). Recovery RL: Safe Reinforcement Learning with Learned Recovery Zones. *IEEE Robotics and Automation Letters*, 6(3).

Xu, X., Grizzle, J.W., Tabuada, P., and Ames, A.D. (2018). Correctness Guarantees for the Composition of Lane Keeping and Adaptive Cruise Control. *IEEE Transactions on Automation Science and Engineering*, 15(3), 1216–1229.

Yang, Z., Wang, Z., and Yan, M. (2021). An Optimization Design of Adaptive Cruise Control System Based on MPC and ADRC. *Actuators 2021, Vol. 10, Page 110*, 10(6), 110.

Yoon, D.D., Ayalew, B., Ivanco, A., and Loiselle, K. (2020). Predictive kinetic energy management for an add-on driver assistance eco-driving of heavy vehicles. *IET Intelligent Transport Systems*, 14(13), 1824–1834.