Safe Reinforcement Learning with Probabilistic Control Barrier Functions for Ramp Merging

Soumith Udatha[1] , **Yiwei Lyu**[1] , **John Dolan**[1]
[1]Carnegie Mellon University, Pittsburgh, United States
{sudatha,yiweilyu,jdolan}@andrew.cmu.edu,

Abstract

Prior work has looked at applying reinforcement learning and imitation learning approaches to autonomous driving scenarios, but either the safety or the efficiency of the algorithm is compromised.With the use of control barrier functions embedded into the reinforcement learning policy, we arrive at safe policies to optimize the performance of the autonomous driving vehicle. However, control barrier functions need a good approximation of the model of the car. We use probabilistic control barrier functions as an estimate of the model uncertainty. The algorithm is implemented as an online version in CARLA[12] Simulator and as an offline version on a dataset extracted from NGSIM Database. The proposed algorithm is not just a safe ramp merging algorithm, but a safe autonomous driving algorithm applied to address ramp merging on highways.

1 Introduction

It is a well-known fact that real-word dynamics are more complicated to be modeled in an analytical form in simulators[1]. This disparity between the real dynamics and the approximations involved becomes more prominent when dealing with safety-critical systems. To determine safety of a system we need to have a good idea of the world dynamics involved.

The interaction of self-driving cars with other human drivers can be treated as a safety critical system.[22]. In the area of autonomous vehicle control and planning (Gu et al.Ames et al.Luo et al.), safety is always the primary focus. Ramp merging is a typical,relatively simple scenario where autonomous vehicles interact with humans.Even for a simple driving scenario, Freeway on-ramp merging areas are high-risk areas for motor vehicle crashes and conflicts due to the variety of driving styles[14] Human drivers introduce uncertainty into merge scenarios, and Autonomous Vehicles (AV) will collide if they can't plan and be controlled safely[33]. It is therefore important that we address model uncertainty while modeling the control and planning of self-driving cars.

In situations with rule-based systems designed to adhere to various driver behaviors on a highway, strictly following the rules to maintain safety could result in unsafe situations, where the autonomous vehicle might brake for an instant when going very close to a vehicle, which may not be anticipated by the human drivers of the tailgating vehicles, leading to collision [32]. The efficiency and solution feasibility of rule-based systems become even worse [27] in situations with model uncertainty. Hence, we need to take human behavior into account along with model uncertainty for better safety guarantees.

We closely study a highway on-ramp merging scenario, where the goal is to enable the AV to merge with human-driven cars safely and efficiently. Our main contributions are in extending and providing a reinforcement learning (RL) pipeline with probabilistically safe constrained barrier functional safety [21][22][6] which can be extended to any self-driving task in a simulation domain by changing the task-based reward function. We perform experiments in CARLA (Dosovitskiy et al.) and an offline training procedure on the NGSIM dataset for a one-on-one ramp merging case. The emphasis of the paper is not on proposing a novel RL algorithm for autonomous cars, but on addressing the problem of model uncertainty with Control Barrier Functions (CBFs) (Ames et al. 6) combined with a RL pipeline to extend the research on safe RL for both offline and online environments.

2 Related Work

CBF-based methods (Ames et al. 6, Ames et al. 4, Notomista et al. 26, Lindemann et al. 19) are being increasingly used in the control application domain due to their safety guarantees with the forward-invariant property when a solution is feasible. In reality,a solution may not always be feasible as we deal with approximations of complex non-linear dynamics and the systems resort to hard-coded recovery control [30].

To address model uncertainty, several works [26][18] proposed to employ the CBF approach with noisy system dynamics. Integrating CBF with Model Predictive Control (MPC) is also a common planning method. [31] and [28] introduced MPC-based safety-critical control. However, these works fail to take different driving behavior styles into account. To address this problem, one possible solution is to use CBFs with reinforcement learning-based methods, where the agent trains for all the possible kinds of behaviors, given enough training data and even a black-box model. RL-CLF-CBFs proposed in [9] is used to reduce the uncertainty in plant dynamics. [23] implemented an RL algorithm with Constrained Policy Optimization (CPO) [2] and with the added safety of control barrier functions.

Generalized CBFs are introduced in [24] to consider higher-order dynamics, Parametric CBFs [29] and Exponential CBFs [25] are popular methods to model approximate higher-order dynamics. In our work, we integrate Probabilistic CBFs [10][21][22] to model uncertainty owing to its ability to use a simpler dynamics model and linear constraints for the optimization problem compared to other methods. Ma et al. is closely relevant prior work, which uses Generalized Discrete CBFs to generate safety certificates and combine with a Model Based Policy Optimization (MBPO) [17] framework for an intersection scenario. Although we have a similar RL pipeline to Ma et al., ours is not a model-based approach and is similar to a CPO formulation. Further, we also introduced a reinforcement learning approach to CARLA and an offline NGSIM Dataset to show the generalization capabilities of our approach.

3 Problem Formulation

3.1 Setting up the problem and Dynamics

Ramp merging has previously been studied in one dimension, we have extended the approach to two dimensions to include a wider range of ramp-merging scenarios and easy generalization to other scenarios. The goal is to control the ego vehicle on the ramp to merge safely with the human-driven vehicles(host vehicles) already on the highway. The system dynamics of a vehicle can be described by double integrators as follows, since acceleration plays a key role in the safety considerations.

$$\dot{X} = \begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{v_x} \\ \dot{v_y} \end{bmatrix} = \begin{bmatrix} 0_{2\times 2} & I_{2\times 2} \\ 0_{2\times 2} & 0_{2\times 2} \end{bmatrix} \begin{bmatrix} x \\ y \\ v_x \\ v_y \end{bmatrix} + \begin{bmatrix} 0_{2\times 2} & I_{2\times 2} \\ I_{2\times 2} & 0_{2\times 2} \end{bmatrix} \begin{bmatrix} u_x \\ u_y \\ \epsilon_x \\ \epsilon_y \end{bmatrix} \quad (1)$$

where $x \in \mathcal{X} \subset R, y \in \mathcal{Y} \subset R, v \in R^2$ are the position and linear velocity of each car respectively and $u \in R^2$ represents the acceleration control input. $\epsilon \sim \mathcal{N}(\hat{\epsilon}, \Sigma)$ is a random Gaussian variable with known mean $\hat{\epsilon} \in R^2$ and variance $\Sigma \in R^{2\times 2}$, representing the uncertainty in each vehicle's motion.

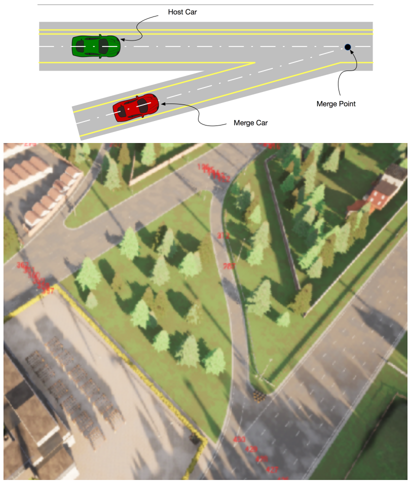

Figure 1: Ramp merging scenario [11]. The ego vehicle (green) is the host vehicle; the merging vehicle (red) is the ego vehicle, running on the ramp.The figure below shows a two dimensional version of ramp-merging that we will be using. The merge is along a curve rather than a straight line.

3.2 Probabilistic Control Barrier Functions

In [22], an optimization framework with probabilistically safe CBFs was proposed as an alternative to deterministic CBFs with perfect model information to address motion uncertainity, specifically for ramp merging scenarios in one dimension.

Consider the admissible space equation for linear CBF with parameter α of one-dimension,x ; $\dot{h}(x,u) + \alpha h(x) \geq 0$ with $h = (x_e - x_m)^2 - R_{safe}^2$. By substituting Eq. 1 we have:

$$\dot{h}_{em}^{s}(x,u) + \alpha h_{em}^{s}(x) \geq 0; \implies 2\Delta x_{em}^{T} \Delta\epsilon_{em} \geq -2\Delta x_{em}^{T}(\Delta v_{em} + u_e \Delta t) - \alpha h_{em}^{s}(x) \quad (2)$$

where $\Delta x_{em} = x_e - x_m, \Delta v_{em} = v_e - v_m, \Delta\epsilon_{em} = \epsilon_e - \epsilon_m \sim \mathcal{N}(\Delta\hat{\epsilon}_{em}, \Delta\Sigma_{em})$ for ego vehicle e and each merging vehicle m. $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function (CDF) of the standard zero-mean Gaussian distribution with unit variance.

We reorganize $\Pr\left(\dot{h}^s_{em}(x,u) + \alpha h^s_{em}(x) \geq 0\right) \geq \eta$ into the form of $\Pr(a^T c \leq b) \geq \eta \Leftrightarrow b - \bar{a}^T c \geq \Phi^{-1}(\eta)||\Sigma^{1/2}c||^2$ from Blackmore et al. with $a = \Delta\epsilon_{em}, c = -2\Delta x_{em}, b = 2\Delta x^T_{em}(\Delta v_{em} + u_e \Delta t) + \alpha h^s_{em}(x)$ and eventually get a constraint; $A_{em}u_x \leq b_{em}, \quad A_{em} \in R^{1\times 2}, b_{em} \in R$
with

$$A_{em} = -2\Delta x^T_{em}\Delta t;\ b_{em} = 2\Delta x^T_{em}(\Delta v_{em} + \Delta\hat{\epsilon}_{em}) + \alpha h^s_{em}(x) - \Phi^{-1}(\eta)\sqrt{\Delta x^T_{em}\Delta\Sigma_{em}\Delta x_{em}} \tag{3}$$

where, Δt is a time unit.We derived linear control constraints for pairwise chance-constrained safety between ego vehicle and each merging vehicle m.The equations in (3) can be extended to y-dimension as well for 2-dimensional decoupled CBFs.For a more detailed derivation, we refer the reader to Lyu et al.. □

3.3 Reinforcement Learning with Probabilistically Safe CBFs

For better safety guarantees, we can use a reinforcement learning policy to compute high level desirable nominal velocity and choose a CBFs control as output. However, it is computationally expensive to solve two different optimization problems while interacting with a simulator.Our constrained RL framework solves it with a single constrained optimization with linear probabilistic safety certificates. Using the optimization framework defined in Achiam et al. for constrained RL problems, we extend the constraints obtained in both dimensions from 3.2 for the total return objective,$\mathcal{G}$ and set of safe states, $\mathcal{C}$ as:

$$\begin{aligned} &\min_{\theta} L_a = E_{a\sim\pi, s\sim C}[\mathcal{G}] \\ s.t \quad & U_{min} \leq u_e \leq U_{max} \\ & J_{c_x} = E_{u_e\sim\pi}[A_{em_x}u_{e_x}] \leq b_{em_x} \\ & J_{c_y} = E_{u_e\sim\pi}[A_{em_y}u_{e_y}] \leq b_{em_y} \\ & D_p(\theta,\theta_k) = \frac{1}{2}\Delta\theta^T H \Delta\theta \leq \delta \end{aligned} \tag{4}$$

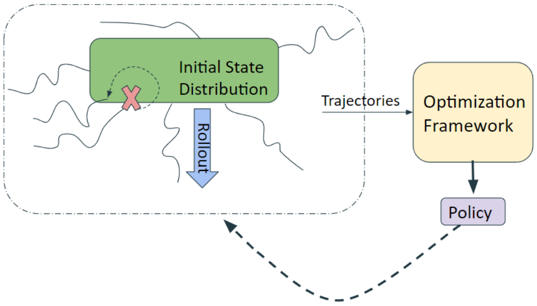

Figure 2: **Reinforcement Learning Pipeline**: We initialize rollouts from random initial states and run an optimization on the set of collected trajectories to train a policy for each epoch.

Using the RL objective in Eq.4, we use approximate actor-critic constrained policy gradient formulation in [23] to solve this optimization. The critic objective is $L(\phi) = E_{s_t\sim\mathcal{C}}[\frac{1}{2}(\mathcal{G} - V(s_t,\phi))^2]$, where ϕ are the parameters of the critic network and $V(s_t,\phi)$ is the critic's state-dependent value function.
The actor objective,(Eq.4) is optimized by transforming into a dual problem and applying KKT (Boyd & Vandenberghe) conditions.The analytical optimal solution is $\Delta\theta^* = \frac{H^{-1}(g - C\nu^*)}{\lambda^*}$ where, θ are actor parameters, λ^*, ν^* are dual variables,H is hessian from trust region condition,$g = \frac{dL_a}{\partial\theta}/||\frac{dL_a}{\partial\theta}||^2$,$C = \frac{dJ}{\partial\theta}/||\frac{dJ}{\partial\theta}||^2$,combining both J_x and J_y.If there is no feasible solution, an iterative line search method using $\theta_{k+1} = \theta_k + -\sqrt{\frac{2*\delta}{C^T H^{-1} C}}H^{-1}C$.A more detailed derivation can be found in the appendix,B. From the RL pipeline in Fig.2, we use multiple rollouts sampled from various initial states, similar to Janner et al. [17], to capture the uncertainty in the environment and to generalize to a larger section of the state space.A pseudocode for our method can be found in Algorithm 1.

4 Experiments

4.1 CARLA

Setting up the Simulation Environment

We use CARLA (Dosovitskiy et al.) v.9.12 to modify the existing Town-06 environment to a ramp merging scenario. The scenario considered in Fig. 1 has a curved ramp. We use a PID Controller to track the steering control of the Ego Vehicle (On-Ramp) between waypoints generated by the default Global Planner to the specified goal position. The control for the Host (Off-Ramp) Vehicle uses both steering and longitudinal PID control. Further, the control of the host vehicle is not reactive. This assumption deals with a conservative case and satisfying this also maintains safety for the case where host control is reactive to the ego vehicle. We optimize the safe reinforcement learning objective (Eq.5) to learn a policy to control the ego-vehicle throttle position.

Algorithm 1 SAPO-RM

```
Require: $\Delta x_{em}, \Delta v_{em}, \Delta t, R_{safe}, \alpha$
Ensure: $\theta, \phi$
  repeat
     Sample trajectories $\mathcal{D} = \mathcal{T} \sim \pi(\theta_k)$
     Compute Constraints in the objective
     Estimate $\phi$ from $\mathcal{D}$
     Calculate g,C,H
     if feasible then
        $\Delta\theta^* = \frac{H^{-1}(g + C\nu^*)}{\lambda^*}$
     else
        $\theta_{k+1} = \theta_k - \sqrt{\frac{2*\delta}{C^T H^{-1} C}}H^{-1}C$
     end if
  until convergence
```

Online Experiments on CARLA

We implemented SAPO-RM in the Simulator set-up in 4.1.1. The RL pipeline introduced in Fig. 2 is followed, where the host vehicle is initialized at a specified initial position with a random offset. If the host vehicle reaches the destination before the ego vehicle, the host vehicle is re-initialized with random offset. The simulation is terminated only when the ego vehicle reaches the destination.The safety distance for the CBFs is 8m. Fig. 3 below shows the results for SAPO-RM for a case where the target host velocity is 40 kph. Training parameters and results for other cases are discussed in the appendix (C).

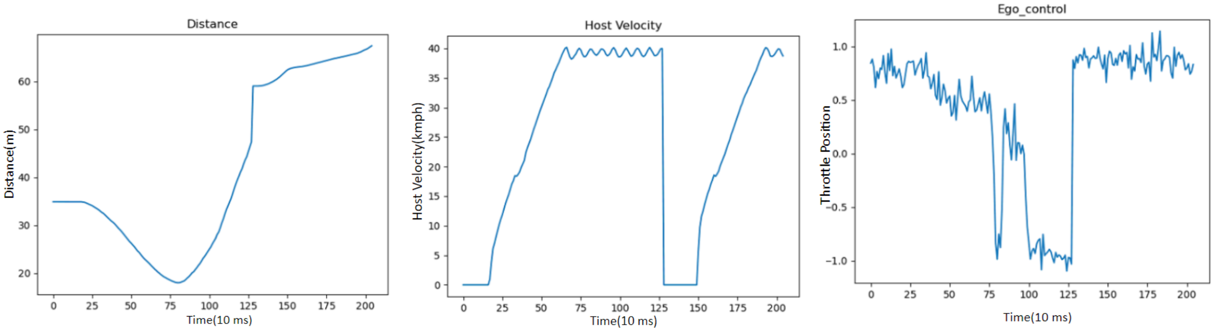

Figure 3: **Results for Online SAPO-RM:** The host velocity is 40 kph, as seen in the center image. The image on the left depicts the distance between the ego and host vehicle.The image on the right is the output of the policy for the evaluated case. From the plots of the ego control (Right) we observe that when distance is close to R_{safe}, the throttle position goes to full-braking, forcing the ego vehicle to maintain the safety condition. Whenever the distance again increases, the ego control goes towards full throttle to optimize performance.

4.2 NGSIM

Data Extraction

An offline dataset for US I-80 highway ramp-merging was extracted from the NGSIM Database. We considered merging vehicles that go from on-ramp to an auxiliary ramp. To ease the complexity, the dataset was simplified into a one-on-one merging. The trajectories are first classified based on the condition: distance between host and the ego vehicle is less than 12m in the auxiliary lane. This condition leads to degenerate values as instances of freeway oscillations are observed in real data (Zheng et al.). We constrained by sorting the data of the ego vehicle with respect to the frame number and considered the first case where the merging condition is satisfied. Further dataset extraction details are in the appendix (D).

Implementing Offline SAPO-RM

The offline dataset obtained in section 4.2.1 is then used to train the offline SAPO-RM policy. Similar to the online case, a rollout is initialized at a random state in the dataset. For each step in a rollout, only the parameters of the ego vehicle are updated according to the dynamics equation (Eq. 1). The vehicle heading direction (ψ) is used as a tracking parameter for the ego vehicle. We used a rollout length of 4 and 5000 trajectories for each epoch of training the offline RL policy for 100 epochs. In Fig. 4, we observe that the black trajectory (Ego policy) deviates from the tracking control to maintain the safe distance.

5 Conclusion and Future Work

We can clearly see from Figs. 3 and 4 that safety conditions are maintained by ego-vehicles while performing ramp-merging.The algorithm we used here is not just a ramp-merging algorithm but is a safe-reinforcement learning algorithm applied to ramp-merging. By relieving some of the assumptions made on the dynamics, the safe-RL algorithm can be extended to other scenarios of lane-changing and intersections. We observe that trajectories maintain the safe distance and are efficient with SAPO-RM. Possible future extensions are extending to situations with more merging vehicles, introducing complicated dynamics models [19], [15] for a better approximation and to alleviate modeling uncertainty, understanding the effect of α on SAPO-RM behavior, and implementing prior methods for the offline approach and other implementations to compare other methods' ability to address modeling uncertainty.

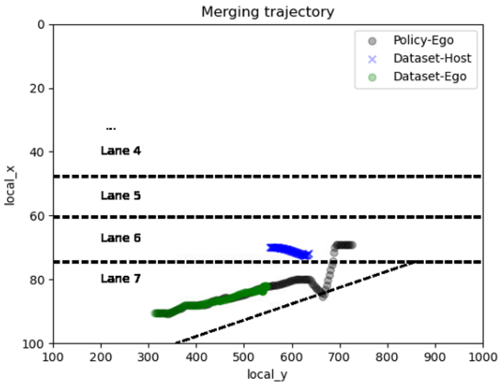

Figure 4: The trajectory in black is generated by a policy, one in blue is the real host vehicle trajectory and the one in green is the real ego vehicle trajectory.

References

[1] AbuAli, N. and Abou-zeid, H. Driver behavior modeling: Developments and future directions. *Int. J. Veh. Technol.*, 2016: 1–12, December 2016.

[2] Achiam, J., Held, D., Tamar, A., and Abbeel, P. Constrained policy optimization. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 22–31. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/achiam17a.html.

[3] Ames, A. D., Grizzle, J. W., and Tabuada, P. Control barrier function based quadratic programs with application to adaptive cruise control. In *53rd IEEE Conference on Decision and Control*, pp. 6271–6278, 2014.

[4] Ames, A. D., Xu, X., Grizzle, J. W., and Tabuada, P. Control barrier function based quadratic programs for safety critical systems. *IEEE Transactions on Automatic Control*, 62(8):3861–3876, 2017.

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

[6] Ames, A. D., Coogan, S., Egerstedt, M., Notomista, G., Sreenath, K., and Tabuada, P. Control barrier functions: Theory and applications, 2019. URL https://arxiv.org/abs/1903.11199.

[7] Blackmore, L., Ono, M., and Williams, B. C. Chance-constrained optimal path planning with obstacles. *IEEE Transactions on Robotics*, 27(6):1080–1094, 2011.

[8] Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, 2004. doi: 10.1017/CBO9780511804441.

[9] Choi, J., Castañeda, F., Tomlin, C. J., and Sreenath, K. Reinforcement learning for safety-critical control under model uncertainty, using control lyapunov functions and control barrier functions, 2020.

[10] Clark, A. Control barrier functions for complete and incomplete information stochastic systems. In *2019 American Control Conference (ACC)*, pp. 2928–2935, 2019. doi: 10.23919/ACC.2019.8814901.

[11] Dong, C., Dolan, J. M., and Litkouhi, B. Intention estimation for ramp merging control in autonomous driving. In *IEEE 28th Intelligent Vehicles Symposium (IV'17)*, pp. 1584 – 1589, June 2017.

[12] Dosovitskiy, A., Ros, G., Codevilla, F., López, A. M., and Koltun, V. CARLA: an open urban driving simulator. *CoRR*, abs/1711.03938, 2017. URL http://arxiv.org/abs/1711.03938.

[13] Duan, J., Liu, Z., Li, S. E., Sun, Q., Jia, Z., and Cheng, B. Adaptive dynamic programming for nonaffine nonlinear optimal control problem with state constraints. *Neurocomputing*, 484:128–141, 2022. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2021.04.134. URL https://www.sciencedirect.com/science/article/pii/S0925231221015848.

[14] El abidine Kherroubi, Z., Aknine, S., and Bacha, R. Novel decision-making strategy for connected and autonomous vehicles in highway on-ramp merging. *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–13, 2021. doi: 10.1109/TITS.2021.3114983.

[15] Ge, Q., Li, S. E., Sun, Q., and Zheng, S. Numerically stable dynamic bicycle model for discrete-time control, 2020. URL https://arxiv.org/abs/2011.09612.

[16] Gu, T., Atwood, J., Dong, C., Dolan, J. M., and Lee, J.-W. Tunable and stable real-time trajectory planning for urban autonomous driving. In *Intelligent Robots and Systems (IROS), IEEE/RSJ International Conference on*, pp. 250–256. IEEE, 2015.

[17] Janner, M., Fu, J., Zhang, M., and Levine, S. When to trust your model: Model-based policy optimization, 2019. URL https://arxiv.org/abs/1906.08253.

[18] Khojasteh, M. J., Dhiman, V., Franceschetti, M., and Atanasov, N. Probabilistic safety constraints for learned high relative degree system dynamics. In *Learning for Dynamics and Control*, pp. 781–792. PMLR, 2020.

[19] Lindemann, L., Robey, A., Jiang, L., Tu, S., and Matni, N. Learning robust output control barrier functions from safe expert demonstrations. *ArXiv*, abs/2111.09971, 2021.

[20] Luo, W., Chakraborty, N., and Sycara, K. Distributed dynamic priority assignment and motion planning for multiple mobile robots with kinodynamic constraints. In *American Control Conference (ACC), 2016*, pp. 148–154. IEEE, 2016.

[21] Luo, W., Sun, W., and Kapoor, A. Multi-robot collision avoidance under uncertainty with probabilistic safety barrier certificates. *Advances in Neural Information Processing Systems*, 33, 2020.

[22] Lyu, Y., Luo, W., and Dolan, J. M. Probabilistic safety-assured adaptive merging control for autonomous vehicles. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10764–10770, 2021. doi: 10.1109/ICRA48506.2021.9561894.

[23] Ma, H., Chen, J., Eben, S., Lin, Z., Guan, Y., Ren, Y., and Zheng, S. Model-based constrained reinforcement learning using generalized control barrier function. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4552–4559, 2021. doi: 10.1109/IROS51168.2021.9636468.

[24] Ma, H., Zhang, X., Li, S. E., Lin, Z., Lyu, Y., and Zheng, S. Feasibility enhancement of constrained receding horizon control using generalized control barrier function. In *2021 4th IEEE International Conference on Industrial Cyber-Physical Systems (ICPS)*, pp. 551–557, 2021. doi: 10.1109/ICPS49255.2021.9468220.

[25] Nguyen, Q. and Sreenath, K. Exponential control barrier functions for enforcing high relative-degree safety-critical constraints. In *2016 American Control Conference (ACC)*, pp. 322–328, 2016. doi: 10.1109/ACC.2016.7524935.

[26] Notomista, G., Wang, M., Schwager, M., and Egerstedt, M. Enhancing game-theoretic autonomous car racing using control barrier functions. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5393–5399, 2020.

[27] Prakken, H. On the problem of making autonomous vehicles conform to traffic law. *Artif. Intell. Law*, 25(3):341–363, September 2017.

[28] Son, T. D. and Nguyen, Q. Safety-critical control for non-affine nonlinear systems with application on autonomous vehicle. In *IEEE 58th Conference on Decision and Control (CDC)*, pp. 7623–7628, 2019.

[29] Taylor, A. J. and Ames, A. D. Adaptive safety with control barrier functions. In *2020 American Control Conference (ACC)*. IEEE, jul 2020. doi: 10.23919/acc45564.2020.9147463. URL https://doi.org/10.23919%2Facc45564.2020.9147463.

[30] Wang, L., Ames, A. D., and Egerstedt, M. Safety barrier certificates for collisions-free multirobot systems. *IEEE Transactions on Robotics*, 33(3):661–674, 2017.

[31] Zeng, J., Zhang, B., and Sreenath, K. Safety-critical model predictive control with discrete-time control barrier function, 2020.

[32] Zheng, Z., Ahn, S., Chen, D., and Laval, J. Freeway traffic oscillations: Microscopic analysis of formations and propagations using wavelet transform. *Procedia Soc. Behav. Sci.*, 17:702–716, 2011.

[33] Zhu, J. and Tasic, I. Safety analysis of freeway on-ramp merging with the presence of autonomous vehicles. *Accid. Anal. Prev.*, 152(105966):105966, March 2021.

A Arriving at Probabilistic Safe Constraints

The chance constrained optimization framework introduced in [22] to handle stochastic world dynamics can be extended into two dimensions, x and y explicitly in our implementation to accommodate uncertainty in the model information.The equation below shows the CBF-QP with a probabilistic chance constraint in one-dimension:

$$\begin{aligned} & \min_{u_e \in \mathcal{U}_e} ||u_e - \bar{u}||^2 \\ s.t \quad & U_{min} \leq u_e \leq U_{max} \\ & \Pr\left(\dot{h}^s_{em}(x,u) + \alpha h^s_{em}(x) \geq 0\right) \geq \eta, \qquad \forall m \end{aligned} \tag{5}$$

where $\bar{u}$ is the nominal expected acceleration for the ego vehicle to follow, and U_{max} and U_{min} are the ego vehicle's maximum and minimum allowed acceleration. $\Pr(\cdot)$ denotes the probability of a condition to be true. $\Pr\left(\dot{h}^s_{em}(x,u) + \alpha h^s_{em}(x) \geq 0\right) \geq \eta \implies \Pr\left(h^s_{em}(x) \geq 0\right) \geq \eta$ given the forward invariance set theory in a deterministic setting, $\dot{h}(x,u) + \alpha h(x) \geq 0 \implies h(x) \geq 0$ as proved in [5] through a chance constraint.

From [7], a general chance constraint problem can be transformed to a deterministic constraint as

$$\begin{aligned} & \Pr(a^T c \leq b) = \Phi(\frac{b - \bar{a}^T c}{\sqrt{c^T \Sigma c}}) \\ \implies & \Pr(a^T c \leq b) \geq \eta \Leftrightarrow b - \bar{a}^T c \geq \Phi^{-1}(\eta)||\Sigma^{1/2} c||^2 \end{aligned} \tag{6}$$

B Constrained Policy Gradient Approximation

The Reinforcement Learning objective of the actor in Eq.4 can be transformed into a dial objective as:

$$\begin{aligned} & \min_{\Delta\theta} g^T \Delta\theta \\ & s.t.\ z + C^T \Delta\theta \leq 0 \\ & \tilde{D_p} \sim \frac{1}{2}\Delta^T H \Delta\theta \leq \delta \end{aligned} \tag{7}$$

where $g = \frac{dL_a}{\partial\theta}/||\frac{dL_a}{\partial\theta}||^2$,$C = \frac{dJ}{\partial\theta}/||\frac{dJ}{\partial\theta}||^2$, $z = J_c - b_m$. Using a Lagrange multiplier, the Lagrangian function then is:

$$\mathbf{L}(\theta, \nu) = g^T \Delta\theta + \lambda(\frac{1}{2}\Delta\theta^T H \Delta\theta - \delta) + \nu(z + C^T \Delta\theta)$$

λ, ν are dual variables. Using KKT Conditions [8],we get the equations:

$$\begin{aligned} & \frac{\partial \mathbf{L}}{\partial \Delta\theta} = g + \Delta\theta + \nu C \\ & \lambda(\frac{1}{2}\Delta\theta^T H \Delta\theta - \delta) = 0 \\ & \nu(z + C^T \Delta\theta) = 0 \\ & \lambda, \nu \geq 0 \\ & (\frac{1}{2}\Delta\theta^T H \Delta\theta - \delta) \leq 0 \\ & (z + C^T \Delta\theta) \leq 0 \end{aligned} \tag{8}$$

solving which, an optimal update direction can be obtained.If there is a feasible solution, the optimal update direction is

$$\Delta\theta^* = \frac{H^{-1}(g - C\nu^*)}{\lambda^*}$$

where ν^*, λ^* are optimal dual variables. For the retrieval update , we ignore the objective function and take the gradient descent with respect to the constraints to force the policy back to the safety region. Similar to [2], $\theta_{k+1} = \theta_k + -\sqrt{\frac{2*\delta}{C^T H^{-1} C}} H^{-1} C$ is the retrieval policy.

The check for feasibility can be done by solving the optimization problem, proposed in [13]:

$$\begin{aligned} & \min_{\Delta\theta} \Delta\theta^T H \Delta\theta \\ & s.t.\ z + C^T \Delta\theta \leq 0 \end{aligned} \tag{9}$$

If the optimal value of the optimization problem above is δ_{min}, the feasible solution set is empty if $\delta_{min} \geq \delta$ and contains a solution otherwise. This problem is optimized efficiently using the Lagrangian Dual Problem:

$$\max_{\nu \geq 0} \frac{-\nu^T C^T H^{-1} C \nu}{2} + \nu^T z$$

A feasibility check can be done by comparing δ_{min} with δ

C Running experiments with CARLA Environments

C.1 Training Parameters and Environment Setup

The state dimension is 37 as we have to reproduce the entire state of an initial state distribution, we save the positions, velocities, angular velocities, throttle and steering positions in three dimensions of both the ego and merging vehicle.The action dimension is 1, as we are only controlling the throttle position. We use 100 trajectories with a rollout length of 20. The actor policy has 2 MLP layers with a size of 256, and the critic is also MLP of size 256. The parameter α is used as 0.75 , the uncertainty parameter η is 0.99 , Δt is 0.01.The reward function we use is negative weighted sum of error from the current position to the merging position chosen on the Town-06 environment map and the error of the velocity to the desired velocity. We were using a desired velocity of 35 kmph for the ego-vehicle. There are no bounds on the throttle position, other than the specified range in Carla from [-1,1].

C.2 Variation with Host Velocity

We saw the results for the host velocity of 40kmph in section 4.1.2. In the figure below, we can see the results with the cases where the host velocities is 30kmph and 20kmph. From the figures 5,6 below it is evident that the minimum distance of 8 is respected with a range of intial conditions that we train our policy on. However, once we step outside the training distribution, we cannot guarantee that this might generalize.

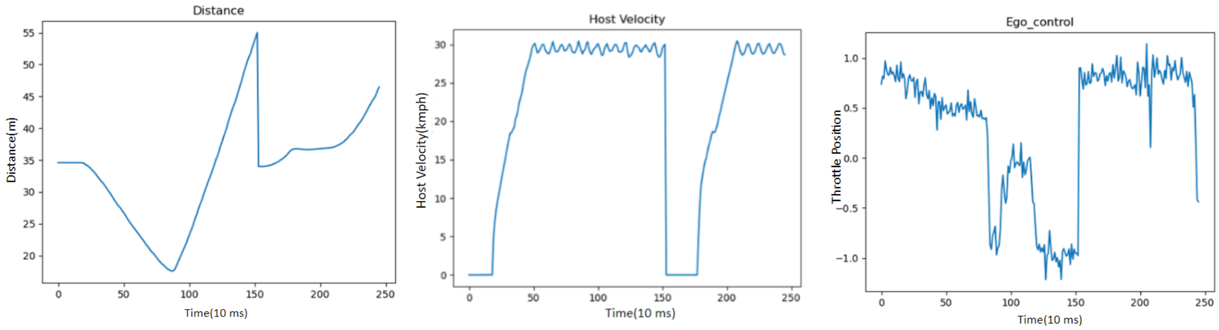

Figure 5: **Host velocity - 30 kmph**

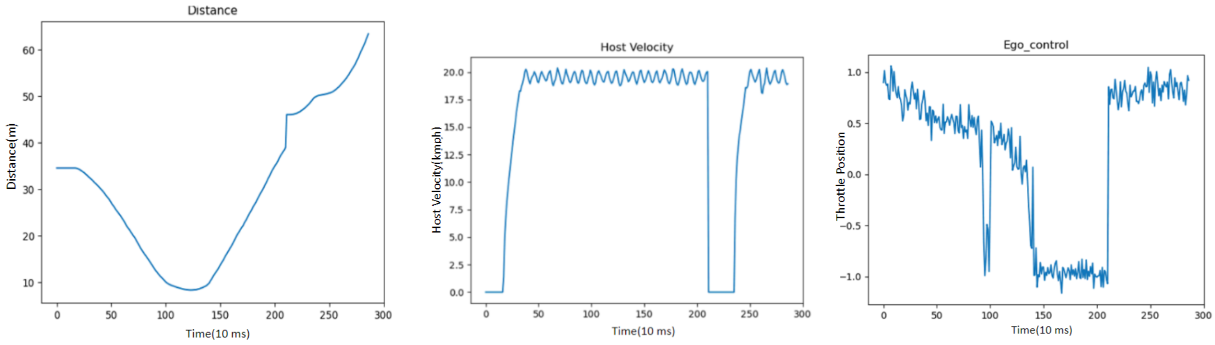

Figure 6: **Host velocity - 20 kmph**

C.3 Variation with α parameter of the CBF

In Figure 7, we repeat the experiment of the ramp merging case with Host velocity 20 kmph as observed in Figure 6. With a higher value of alpha, ideally we should observe more aggressive behavior, which is also the case from the plots in Fig.7. One main difference between the two versions is that the merging agent waits before the merge for $\alpha = 0.75$ and merges behind the host vehicle. For the case with higher α, the merging vehicle accelerates from the beginning and would merge the host vehicle in front of it.

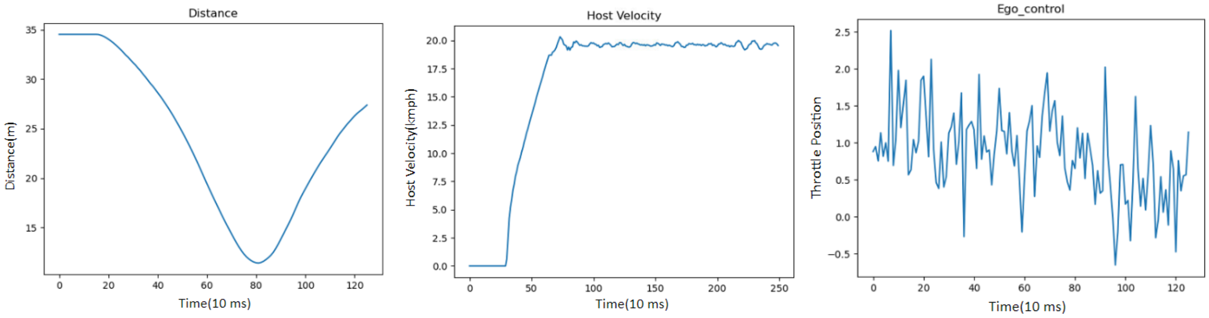

Figure 7: **Host velocity - 20 kmph and α = 15**

D NGSIM Dataset Extraction

Of the possible 11,850,526 data points in the NGSIM Dataset we have extracted a train, validation and test split of the merging dataset of about 100,000 data points of the merging trajectories for an one-on-one situation. The figure below shows the trajectories of the merging condition we have applied on the US I-80 highway dataset. We filter the ego vehicles that start in lane 7. Of the possible combinations to go to lane 5 and lane 6, we only consider the case where vehicles merge to lane 6. For the host vehicles, we consider the vehicles that are within a specified distance of 12, when the ego vehicle is merging to lane 6. We also sort the data in terms of frame numbers of ego vehicles for time-based trajectories.The trajectories generated can be seen in Figure 7. The network parameters of actor and critic, and the parameters of α, η andΔt are the same as that for the online case. The reward function in this case is negative tracking error between the current trajectory and the dataset trajectory for ego-vehicle. More results from different initial positions are shown in figure 8.We'll report more results on the average increase in distance and time performance gain on the dataset in our future work or in the camera ready version.

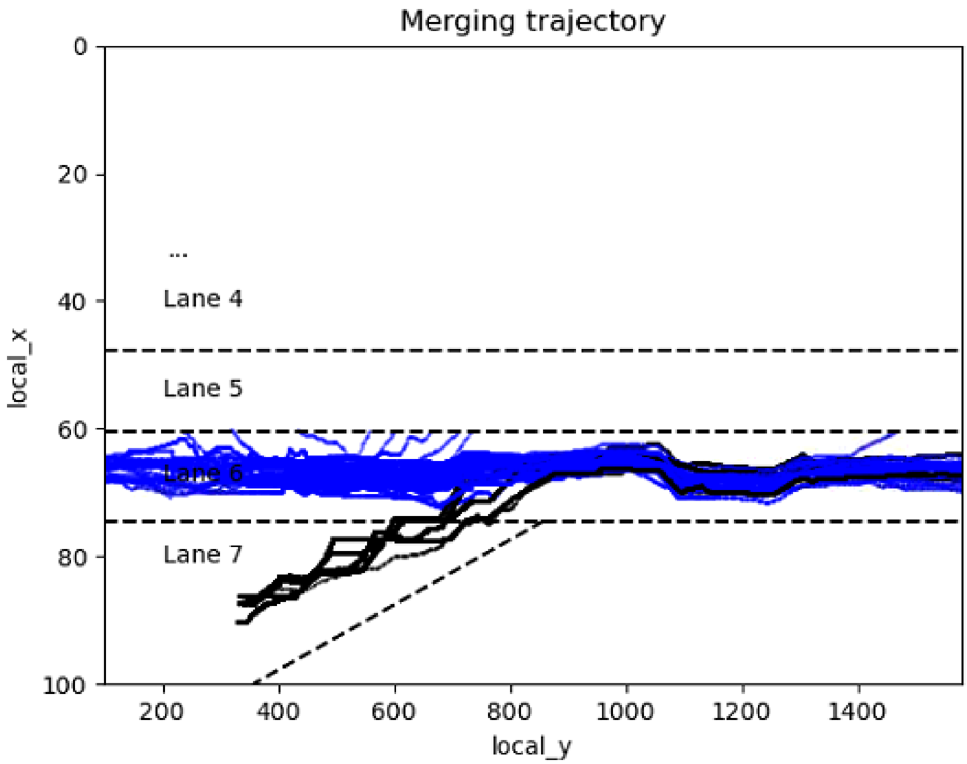

Figure 8: **Merging Dataset:** It depicts all the trajectories extracted for I-80 dataset for one-on-one ramp merging.Trajectories in blue are that of the host vehicle and in black are of the ego vehicle.

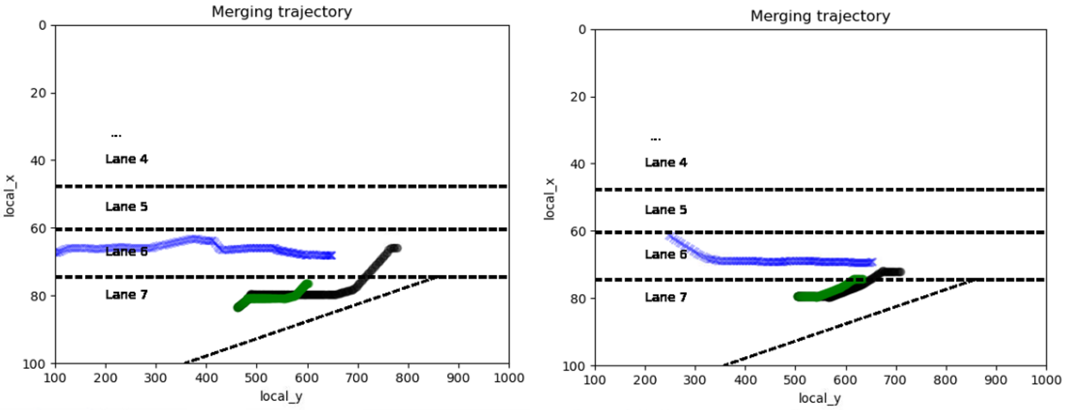

Figure 9: **Offline Sapo RM:** The results trajectories shown here compare the green and black trajectories, which are the orignal trajectory in the dataset and the RL policy trajectory respectively.

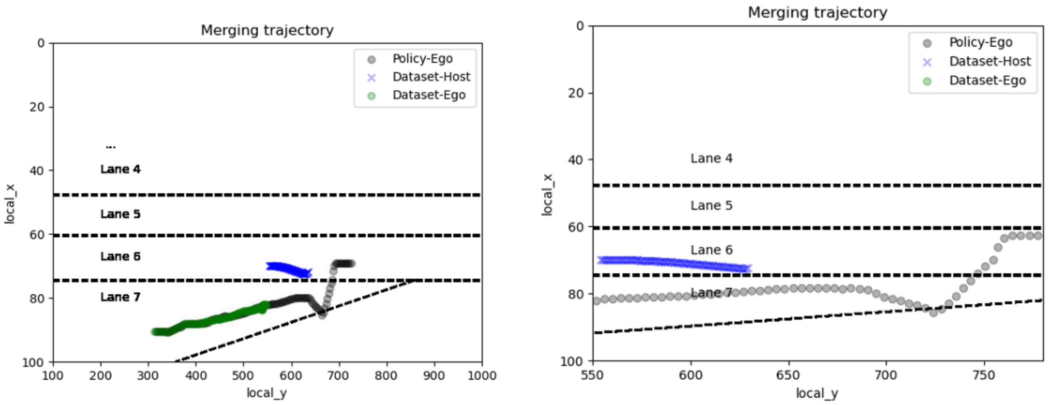

Figure 10: **Observing Trajectory:** On the image to the left we observe the entire trajectory of Fig.4 and on the figure on the right we looked at a zoomed in version to observe the same trajectory in more detail.As the figure on the left has a different scale, it appears that the trajectories are very sharp but with a detailed trajectory on the right, the change in the heading direction of the car is not very sharp but rather an illusion of scale.