Learning Safe Driving with CVaR Fine Tuning

Kusha Sareen
School of Computer Science
McGill University
kushagra.sareen@mail.mcgill.ca

Abstract

A key step in the deployment of autonomous driving agents is to ensure their safety, especially in the worst case. The Conditional Value at Risk (CVaR) measures the return in these worst-case outcomes. This work trains two algorithms: the Advantage Actor Critic (A2C) and Deep Q Network (DQN) in the `highway` environment. Experiments show that lowest quartile returns can be drastically improved by fine tuning a model to optimize the CVaR of the return over a batch of trajectories.

1 Introduction

Ensuring the safety of autonomous driving agents is essential before deployment. However, when learning to drive, there is a trade-off between learning cautious driving practices and the flexibility of an agent that looks, for instance, to optimize its speed. Ideally, we prefer agents that are capable of complex maneuvers in dire situations but choose to drive safely and conservatively the majority of the time. How can we train an agent to learn driving best-practices without losing out on maneuverability?

For instance, penalizing an agent too heavily for crashes hinders its exploration, meaning that the agent may never learn more complex movement while rewarding speed and movement too much creates unsafe drivers. This work proposes an alternative solution: train the agent to optimize movement then fine tune to promote safe driving. The intuition is that complex movement is a skill to be learned, whereas safe driving practices are more like a preference that can be easily adapted.

2 Background

2.1 Environment

This work is an exploration of the `highway` environment, made by the Farama Foundation [2] to learn tactical decision making while driving . The environment consists of a controlled car driving on a 3-lane highway with other autonomous vehicles (Figure 1). The agent has 5 meta-actions it can take at any timestep: idle, faster, slower, merge left and merge right. The other vehicles are controlled according to the Intelligent Driver Model from [5] with realistic properties. Agent reward is given by

$$R(s, a) = a\frac{v - v_{min}}{v_{max} - v_{min}} - b * collision,$$

where v_{min} and v_{max} are the minimum and maximum allowed speeds on the highway respectively and a, b are parameters. An agent's observation is given by tuples $(presence, x, y, v_x, v_y)$ for itself and the 4 cars in its immediate proximity, where $presence$ defines whether such a car exists, and $(x, y), (v_x, v_y)$ are normalized position and velocity vectors respectively.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

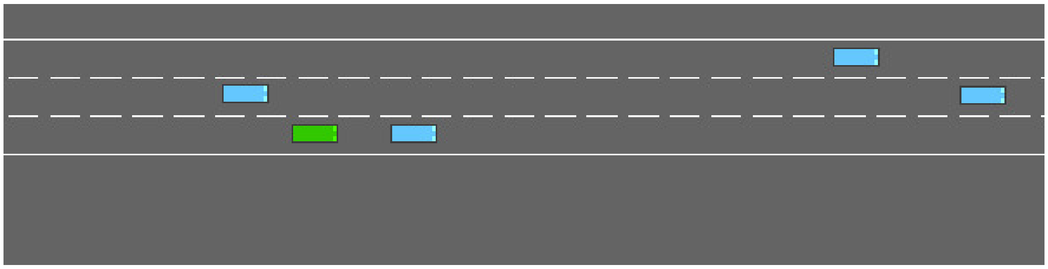

Figure 1: The `highway` environment configured with 3 lanes and traffic density 1. The green vehicle is the agent and blue vehicles are controlled by the Intelligent Driver Model.

2.2 Models

Deep Q Network (DQN) and Advantage Actor Critic (A2C) models were trained in the environment. The DQN uses a function approximator to learn a state-action value function $Q_\theta(s, a)$. With some exploration, actions are chosen according to $Q_\theta(s, a)$. Over training, we optimize θ such that

$$L_{DQN}(\theta) = \mathbb{E}_{(s,a,r,s')\sim\mathbf{D}}[(r + \gamma \max_{a'} Q_\theta(s', a') - Q_\theta(s, a))]$$

is minimized over the collected data $\mathbf{D}$ with horizon γ.

The A2C trains a policy $\pi_\theta(s)$ and a state value function $V_w(s)$ such that

$$L_{A2C}(\theta, w) = \mathbb{E}_{(s,a,r,s')\sim\mathbf{D}}[-\log \pi_\theta(a|s)(R_t - V_w(s)) + c(R_t - V_w(s))^2]$$

is minimized over training, where $R_t = \sum_{t'=t}^{T-1} \gamma^{t'-t} r_{t'} + \gamma^{T-t} V_s(s_T)$ is an estimation of the return over the remainder of the episode and c is a parameter. The reader can refer to [4] and [3] for an in-depth discussion of each of the models.

2.3 CVaR Tuning

The Conditional Value at Risk (CVaR) is a generalization bridging the gap between the mean and minimum of some distribution. For a distribution with CDF $F_X(x)$ and α-quantile $q_\alpha(X) = \inf\{x|F_X(x) \geq \alpha\}$, the CVaR is defined by

$$CVaR_\alpha(X) = \mathbb{E}[x|x \leq q_\alpha(X)].$$

By optimizing the CVaR, we modify the training objective to focus on the tail of the return distribution. In effect, this shifts the baseline in our gradient from the mean return $\bar{G}$ to $CVaR_\alpha(G)$. As a result, behaviours leading to irregularly low returns (crashing, for instance) are discouraged, perhaps at the cost of lower total returns.

Using CVaR as a training objective in reinforcement learning has been previously studied (see [6]) and inspiration for this work was taken from [1] where it is used in a meta-learning setting.

3 Methodology

Each of the models were trained for 400 episodes in the environment for 5 runs after heuristic optimization of hyper-parameters.

The chosen model architectures are as follows. The A2C neural network passes a length 25 observation vector into a common layer with 50 neurons and outputs probabilities for each of the 5 actions and a state value. Softmax activation is used for the probabilities and ReLU is used everywhere else. A learning rate of 0.001 was chosen with parameter $c = 1$.

Similarly, the DQN network takes the same state, has a common layer with 50 ReLU neurons and outputs values for each action. A learning rate of 0.01 was chosen. Actions are selected with

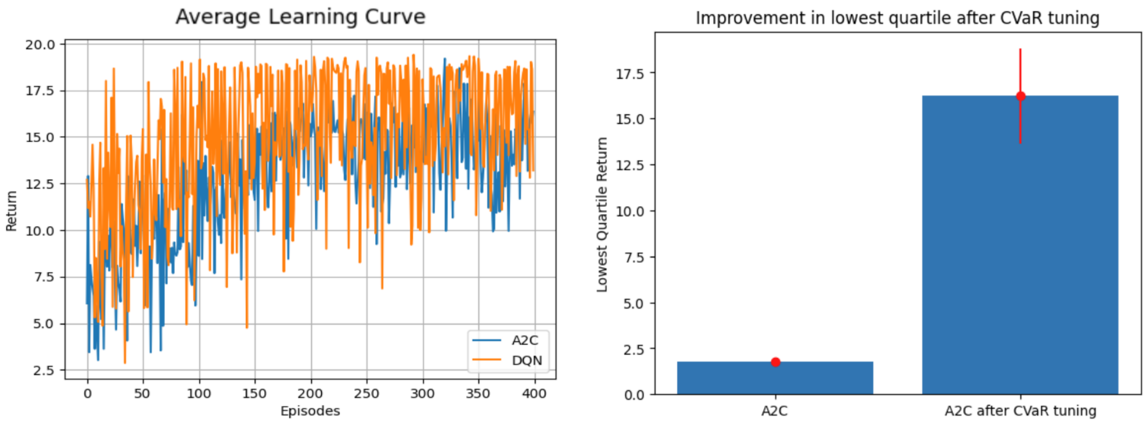

Figure 2: (Left) Average learning curves depicting returns for the DQN and A2C for 5 runs over 400 episodes. (Right) Lowest quartile returns for A2C over 100 episodes of testing, before and after fine tuning to optimize the CVaR.

Boltzmann exploration with temperature $T = 1$. Both models were trained with a horizon of $\gamma = 0.99$, with gradients applied at the end of each episode with the Adam optimizer.

The best A2C model was selected for fine tuning. Fine tuning was performed over 100 batches. For each batch, K episodes of data are collected, and gradients are applied from episodes with lowest $\alpha\%$ of returns in a given batch. This gives an approximation of the CVaR. Gradients are applied every $K = 10$ episodes and we pick $\alpha = 25\%$ of the lowest returns. Lowest quartile returns are plotted for the same A2C agent over 100 episodes, before and after fine tuning.

All computations were performed on the Google Colab servers with NVIDIA Tesla V100 GPU.

4 Experimental Results

Plots of learning curves and lowest quartile returns before and after tuning are given in Figure 2. Additionally, videos of each agent are available in the supplementary material.

Learning curves have high variance largely because of the variability of training time and the variable nature of the environment. Some agents were able to a learn a good policy in $\sim$100 episodes and received returns $\sim$20 for the remainder of the training period, albeit with some variance, whereas others did not learn a good policy over the entirety of training and maintained returns $\sim$0. This difficultly likely stems from the variable nature of the environment. Other cars on the highway respond to the behaviour of the agent and drive accordingly. Thus, the distribution of other cars' actions an agent observes is contingent on its own actions, where an agent that hasn't learned good behaviours postpones seeing how other cars would behave in response.

Overall, we observe similar final performance between the best DQN and A2C agents. Indeed, videos show the best agents learn a similar policy. The agents are rewarded for their speed and so rush through traffic. After fine tuning, the agent is much more conservative and prefers to stay in its lane but it seems dexterity is maintained. Indeed, we can see the lowest quartile returns are drastically increased after CVaR tuning. This likely eliminated a behaviour, for example immediately switching lanes to overtake, that would sometimes lead to irregularly low returns. After tuning, lowest quartile returns better reflect the average performance of the best agents.

5 Conclusion and Future Work

In conclusion, this experiment hopes to demonstrate the effectiveness of optimizing the CVaR metric in driving agents to enhance safety performance. After fine tuning, the agent was able to effectively reduce the likelihood of dangerous driving behaviors while still maintaining efficient and smooth performance. This provides a promising solution for enhancing the safety of autonomous driving systems while maintaining their dexterity.

Future work can build upon these results to further improve the safety and reliability of the driving agent. One possible avenue is to explore impact reward functions on learned behaviour. For instance, can a comparable policy be learned without fine tuning by choosing a particular reward function? Additionally, future work can aim to make driving style more generally adaptable. After learning to drive, techniques such as meta-learning, can be to enhance the agent's ability to adapt to driving scenarios and environments the particular agent is more likely to encounter, for instance regional differences in driving culture and practices. This was explored in this project but without any affirmative results.

Finally, the effectiveness of the proposed CVaR optimization approach should be evaluated under more diverse driving conditions and scenarios, such as adverse weather, variable roads, or high-density traffic environments. Overall, these findings highlight the potential of CVaR fine tuning to ensure the safe operation of autonomous driving agents.

References

[1] I. Greenberg, S. Mannor, G. Chechik, and E. Meirom. Train hard, fight easy: Robust meta reinforcement learning, 2023.

[2] E. Leurent. An environment for autonomous driving decision-making. `https://github.com/eleurent/highway-env`, 2018.

[3] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu. Asynchronous methods for deep reinforcement learning. *CoRR*, abs/1602.01783, 2016.

[4] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. A. Riedmiller. Playing atari with deep reinforcement learning. *CoRR*, abs/1312.5602, 2013.

[5] M. Treiber, A. Hennecke, and D. Helbing. Congested traffic states in empirical observations and microscopic simulations. *Physical Review E*, 62(2):1805–1824, 2000.

[6] C. Ying, X. Zhou, H. Su, D. Yan, N. Chen, and J. Zhu. Towards safe reinforcement learning via constraining conditional value-at-risk, 2022.