# Finding the Value of Aggressiveness in Autonomous Vehicles

William Brannon (wbrannon)

December 2019

**Abstract**

This project involves modeling a lane-changing process of an autonomous vehicle among simulated human-driven vehicles, in which the autonomous vehicle learns a state-action mapping via deep reinforcement learning. The problem was modeled as a Markov Decision Process (MDP) with a discretized action space built by variations in lateral and longitudinal acceleration. The objective in which this project is defined revolves around finding the value of incorporating "aggressiveness" in an autonomous vehicle among human drivers. Initial results were gathered in which the autonomous vehicle was made to perform the maneuver among five human-driven vehicles on a four-lane straight roadway, and final results involved a more densely populated environment among 40 human-driven vehicles. It was found that this approach yielded safer results for an ego agent exhibiting less aggressive behavioral characteristics.

## 1 Introduction

There exists a heavy amount of work currently done regarding decision making and planning for autonomous vehicles, of which the corresponding projects often incorporate "human drivers," represented by a simulated vehicle operated by a set of given parameters. However, much recent work fails to capture the influence that an autonomous vehicle's actions may have on a human, and therefore some of the most effective interactions between the two are often missed; for example, in April 2018 a Waymo vehicle was shown to fail a merge onto a freeway due to its characteristics as a "passive" driver. Due to problems such as this, it becomes necessary to explore methods of capturing the more "aggressive" behavior that is often displayed by humans.

There is a large variety of potential inputs for a problem such as this, but this application utilizes x and y coordinates along with velocities for each vehicle in the scene to input; due to simplifications made in the simulation process (explained in section 3), this is sufficient for input data. Ultimately, this project seeks to find the value in incorporating a higher "aggression" level in an autonomous vehicle among human drivers in a lane-changing scene; it is assumed that this "aggression" level is correlated with general trends in high acceleration levels, along with less required space in front of another vehicle during a lane change maneuver.

In modeling the problem as a Markov Decision Process, this process can be effectively explored via deep reinforcement learning. A Markov Decision Process (MDP) is defined by a tuple $(S, A, P, R)$ where $S$ is the state space, $A$ is the action space, $P$ is the transition model, and $R$ is the reward function. Actions are selected via a policy $\pi$, which is governed by a state-action value function $Q(s, a)$, which represents the collected reward at a given state after taking a given action [5]. This project seeks to uncover an optimal Q function given two differing "aggression" levels for the ego agent, and compare the two in terms of amount of collisions caused and the total amount of time taken to get to the end lane; it is theorized that the more aggressive vehicle will perform better in the total amount of time taken to reach the goal lane, while there is less insight into the potential difference in collisions between the two ego agents.

As is done in much recent work [2][3][4], the human vehicles during this experiment are controlled via the Intelligent Driver Model (IDM), which maps features such as distance to the ahead vehicle to a longitudinal acceleration[1].

## 2 Related Work

There exists a heavy amount of literature regarding decision making and planning for autonomous vehicles; however, most projects have not taken into account the influence that the autonomous vehicle may have on the human drivers. This may be accounted for in a vehicle's "aggression" parameters[3][4]. Urban scenarios are a topic of heavy interest, in which the autonomous vehicle is mapped around numerous human drivers. In addition, deep reinforcement learning is heavily used to operate autonomous vehicles [4][7][8]. There also exists work in modeling human drivers in urban scenarios [1]. This work serves as an extension of [3], but assumes full observability and takes into account the aggressiveness of the autonomous vehicle as opposed to that of the human driver.

## 3 Dataset and Features

This project made use of the Stanford Intelligent System Laboratory's driving simulator via **AutomotiveDrivingModels.jl**. This simulator allows an approach to capturing high level vehicle dynamics among a virtual roadway; the roadway may be inhabited by a chosen number of vehicles. Figure 1 displays a snapshot of a simulation. The selected features into the Q network are every vehicle's x and y coordinates and velocity, in which the entire feature vector was normalized. This totals to an amount of input features equal to three times the amount of present vehicles.

One million total samples were fed into the Q network, in which the amount

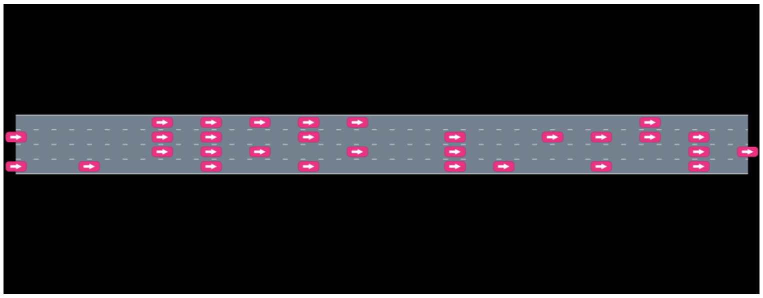

Figure 1: Beginning of a simulation of a lane-changing scenario. 40 human-driven vehicles are pictured, in which the x and y coordinates and velocity are used as inputs to the feature vector. Note that human driven vehicles were distributed randomly on a 150-meter roadway in buckets every 10 meters for simplicity.

of samples (or steps) per episode ranged on the amount of vehicles present in the scene (it was assumed that less steps should be granted in scenes containing less obstacles). At the start of each episode, the ego agent begins in the rightmost lane and is tasked with moving to the leftmost lane. The reward function was dependent upon events such as collisions, reaching the goal state, having an overly large heading angle, and running offroad; a high penalty is assigned to colliding with another vehicle, while a high reward is assigned to reaching the goal lane. The terminal state is dependent on reaching the goal state, colliding with another vehicle, reaching the end of the road, or timing out.

At each timestep, the ego vehicle is given a choice from a discretized set of actions, revolving around speeding up or slowing down, along with moving to the left or right lane, for a total of nine actions; the action space is defined by `(SlowLeft, NormalLeft, SpeedLeft, SlowStraight, NormalStraight, SpeedStraight, SlowRight, NormalRight, SpeedRight)`. These are initially selected at random every 0.1 seconds, before the Q function gains valuable insight into state action values following Bellman optimality.

# 4 Methods

A deep Q network is formulated to estimate the state-action value at each state, in which the optimal Q function is known to follow the Bellman equation:

$$Q^*(s,a) = r \ + \ \gamma_{max'_a}(Q(s',a'))$$

In developing a deep Q network, it is interesting to note that the network output $y$ and target $\hat{y}$ are both included in the optimization process as $Q^*(s,a)$ and $Q(s',a')$, respectively. While $\hat{y}$ is held constant for a given amount of iterations, it is enforced that the optimal Q function follows Bellman optimality, and $y$ is consistently updated accordingly. The loss function of a deep Q network is defined as

$$l(\theta) = \mathbb{E}_{s'}\big[\big(r \ + \ \gamma \max_{a'} Q(s', a'; \theta) - Q(s, a; \theta)\big)^2\big]$$

where $r$ is the reward received at state $s$ after taking action $a$, in which the network is parameterized by $\theta$ [6]. As seen the loss function is governed by the difference between the current Q function and the Q function that follows Bellman optimality. The parameters $\theta$ are updated after receiving an experience same ($s$, $a$, $r$, $s'$) according to the following trend:

$$\theta := \theta \ + \ \alpha\big(r \ + \ \gamma \max_{a'} \big(Q(s', a'; \theta) - Q(s, a; \theta)\big)\nabla_\theta Q(s, a; \theta)$$

# 5 Experiments/Results/Discussion

Two separate ego agents, distinguished by a difference in acceleration parameters (one would naturally gain speed faster than the other) and the amount of space needed to move over into the next lane were compared to find out the difference in safety (governed by the amount of collisions per number of simulations) and efficiency (defined by minimal time to reach the goal lane). To emphasize these objectives, a reward of 1000 was granted to reaching the goal lane, while a penalty of 500 was placed for taking part in a collision. The ego agents were tested in environments containing 5 human agents, 20 human agents, and 40 human agents, with the expectation that performance would degrade with additional agents in simulation.

It was found that over one million steps per scenario, a learning rate of 0.0005, and a target ($Q(s', a')$) update frequency of 10,000 steps yielded highest-performing results. In addition, the advent of prioritized replay yielded positive results on the loss. After training, each agent was placed in 5000 simulations, and interesting results were found in the final performance, as seen in the table below:

| Agent | Amount of Human Vehicles | Collisions | Goal |
|---|---|---|---|
| Passive | 5 | 418 | 3581 |
| Passive | 20 | 1640 | 2360 |
| Passive | 40 | 3800 | 1200 |
| Aggressive | 5 | 355 | 3284 |
| Aggressive | 20 | 2013 | 26 |
| Aggressive | 40 | 2430 | 0 |

As seen in the table, the ego agent characterized by passive parameters outperformed the ego agent characterized by aggressive parameters when surrounded by 20 and 40 human drivers, and slightly underperformed in an environment with five human drivers. This makes intuitive sense, as it can be deemed safer to exhibit more aggressive characteristics when driving in a more spacious environment. Transfer learning was explored with the previously trained weights for both agents with a modification in the reward function to prevent collisions, but similar results were found after several million more experiments. Figure 2 displays the reward curve during training, and it should be noted that half of the training phase is dedicated to exploration via an epsilon greedy strategy.

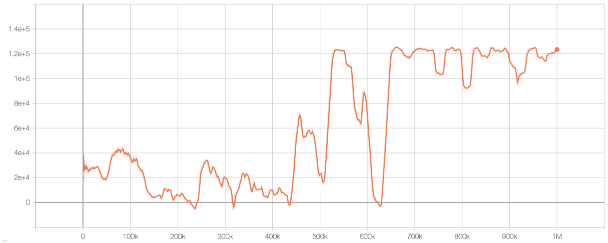

Figure 2: Sample of reward curve during training, with 1/2 of the training dedicated to exploration

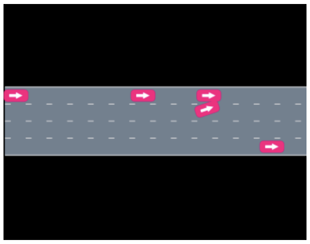

Figure 3: Snapshot of collision in simulation

# 6 Conclusion/Future Work

In terms of future work, it would be necessary to introduce partial observability and formulate the problem as a Partially Observable Markov Decision Process (POMDP) in order to make the scene more realistic to what is seen in real life. For this, one may introduce a belief state on the parameters of the IDM controlled human drivers, in which IDM may be augmented to produce more aggressive or passive drivers. In addition, providing an action space more realistic to actual continuous space would possibly provide for more realism in simulation. In addition, placing a safety checker on the program would prove a safer system and cause less collisions.

Further, because this approach involved placing an untrained agent in a difficult-to-engage environment, better results may be found by starting training on a simpler road, and progressively using transfer learning to move to more difficult environments. The code on this project may be found at https://github.com/wbrannon/CS230-Project.

# 7 References

## References

[1] M. Treiber, A. Hennecke, and D. Helbing, "Congested traffic states in empirical observations and microscopic simulations," *Physical Review E*, vol. 62, no. 2, p. 1805, 2000.

[2] M. Bouton, A. Cosgun, and M.J. Kochenderfer, "Belief State Planning for Autonomously Navigating Urban Intersections," 2017.

[3] Z.N. Sunberg, C.J. Ho, and M.J. Kochenderfer, "The value of inferring the internal state of traffic participants for autonomous freeway driving," in *American Control Conference (ACC)*, 2017.

[4] M. Bouton, A. Nakhaei, K. Fujimara, and M.J. Kochenderfer, "Cooperation-Aware Reinforcement Learning for Merging in Dense Traffic," 2019

[5] M.J. Kochenderfer, *Decision Making Under Uncertainty: Theory and Application.* MIT Press, 2015.

[6] R. Bellman, *Dynamic Programming.* Princeton University Press, 1957.

[7] H. Chae et. al, "Autonomous Braking System via Deep Reinforcement Learning," in *IEEE International Conference on Intelligent Transporation Systems (ITSC*, 2018.

[8] T. Tram et. al, "Learning negotiating behavior between cars in intersections using deep q-learning". Princeton University Press, 1957.