Learning Reward Functions for Optimal Highway Merging

Elliot Weiss

Abstract

In this paper, we explore highway merging for autonomous vehicles as a Markov decision process (MDP). Two different reward function formulations are studied, one which encodes safety and mobility objectives and the other which assumes a simple polynomial form based on state and action values at each time step. Solving this MDP for both reward functions with a variety of transition dynamics prescribing faster or slower driving behavior for another vehicle on the road enables a comparison of Pareto optimal curves for each reward function over an iterative sweep of reward function weights. We find that the assumption-free, polynomial reward function significantly outperforms the reward function that encodes prior knowledge for both safety and mobility objectives. A formal inverse reinforcement learning (IRL) approach to this problem may enable a more rigorous computation of reward functions that best explain optimal merging policies.

1 Introduction

Autonomous vehicles must drive safely and efficiently within stochastic environments. In this paper, we explore the specific problem of merging onto a highway containing other drivers with unknown patterns of behavior. Highway merging is a scenario of great interest in autonomous driving and motion planning research given the complex tensions which exist among multiple agents with competing goals. The autonomous vehicle must optimize multiple objectives, which can be summarized as the need to simultaneously achieve reliable safety levels and move through the environment quickly. At each point in time, the autonomous vehicle can move forward various amounts within the on-ramp to generate a safe gap between it and an adjacent vehicle or merge onto the highway. Optimal driving policies are generated through the value iteration algorithm based on a variety of driving behaviors for other vehicles on the road and simulated within a stochastic environment to determine sets of ideal driving policies based on various reward function formulations.

This paper borrows ideas from inverse reinforcement learning (IRL), which is the process of determining the reward function an agent is optimizing given observations of optimal behavior [1]. IRL research is applied in many domains including economics, behavioral psychology, control theory, and human-centered design. A central motivation behind IRL is the idea that certain tasks – such as cooking a three course meal or determining if another agent is behaving morally – are currently much better executed by human beings than artificial intelligence systems and, therefore, it is useful to explore humans' behavior in these tasks to learn the underlying structure of the rewards they are maximizing. Data of human drivers optimally navigating a complex environment can be utilized to calculate the reward function, which itself is then used in a traditional reinforcement learning setting to determine optimal polices for an autonomous vehicle in new environments. Various IRL algorithms determine reward functions based on knowledge of either the entire optimal policy or a set of observed trajectories that represent a subsection of the full policy [2]. Our approach looks at two potential reward function formulations and, by varying weights within these functions, determines sets of optimal behaviors to compare how well various reward function formulations optimize for quantities of interest.

2 Related Work

There is a wide breadth of previous work on optimal policies for vehicle merging, many of which formulate this problem as a partially observable Markov decision process (POMDP). The task of interacting with pedestrians represents a comparable problem of high importance in autonomous driving research. Online solution methods for driving through a highly dynamic crowd of people [3] and offline methods for incorporating human values into driving behavior at an occluded crosswalk [4] have been explored through a POMDP framework. Further, optimal path planning through an urban intersection has been investigated with a continuous state space formulation, incorporating both offline and online solution methods [5] [6]. Additional research on highway merging has studied techniques for inferring the underlying intentions of other vehicles through a POMDP structure [7] and one-dimensional driving behavior as a Markov decision process (MDP) for robust driving under significant perceptual constraints [8]. Our approach builds on POMDP and MDP problem formulations for various autonomous driving scenarios and seeks policies that balance safety and mobility objectives.

3 Problem Formulation

We model this problem as a discrete state and action space MDP to enable fast computations, particularly given many iterations over reward function weights that we compute. Although autonomous vehicles with imperfect sensors operate with state uncertainty given observations of their external environment, we assume full observability to focus on the relationship between optimal driving strategies and reward function structures. An MDP is defined by the tuple $\{S, A, T, R, \gamma\}$, where S is the state space, A is the action space, T is the transition function, R is the reward function, and γ is the discount factor, which we set to a value of 0.9 for all calculations to balance the value of present and future rewards. The following subsections describe each component of this MDP.

3.1 State Space

The state at each point in time is a three value vector comprised of the ego vehicle's lateral position x_e and longitudinal position y_e and the longitudinal position of a vehicle in the other lane of the highway into which the ego vehicle is attempting to merge y_1. The highway is discretized into 50 longitudinal positions and 2 lateral positions (the on-ramp is lane 1, and the right lane of the highway is lane 2), resulting in a state space of size $|S| = 2 \times 50 \times 50 = 5,000$. There are several terminal states, at which point the simulation ends and no more rewards can be attained. These are all states for which the vehicle has merged ($x_e = 2$) or either vehicle has reached the end of the road ($y_e = 50$ or $y_1 = 50$). Figure 1 shows an example state within this discretized space, indicating the direction of longitudinal vehicle motion.

3.2 Action Space

The action space comprises four actions, represented as $A = \{1, 2, 3, 4\}$. Actions 1, 2, and 3 result in the ego vehicle moving in the longitudinal direction 1, 2, or 3 spaces, respectively. Action 4 results in the ego vehicle merging, which transitions the value x_e from 1 to 2 and increases its longitudinal value y_e by 1 space. It is important to note that taking action 4 always results in a transition into a terminal state.

3.3 Transition Function

The transition function $T(s'|s, a)$ computes the probability of reaching state s' from state s given action a. The transition dynamics for the ego vehicle are deterministic; taking actions 1, 2, and 3 always result in those respective increments in y_e value, and action 4 always executes the same merging maneuver. Although robotic actuation can contain some noise, this assumption is relatively accurate for the specific case of an autonomous vehicle and enables simplicity in our computations. The other vehicle transitions stochastically according to three models of behavior – a fast driving model in which it moves longitudinally 2 steps with a probability of 0.7 and 1 step with a probability of 0.3, an average speed driving model where these probabilities are 0.5 and 0.5, and a slow driving

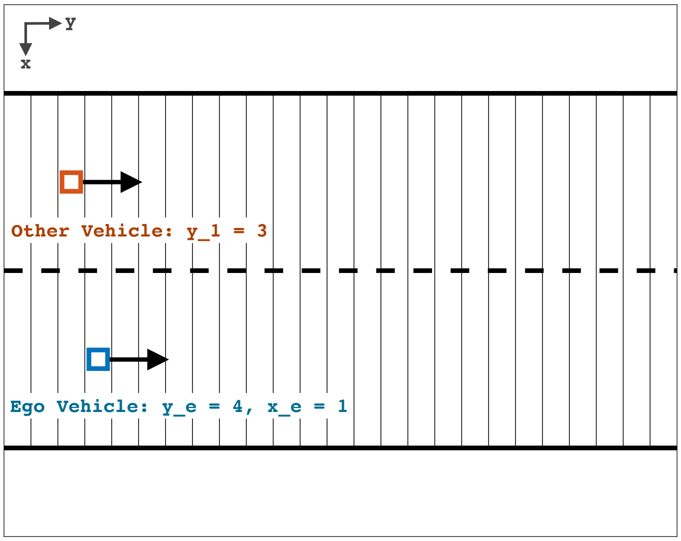


Figure 1: Diagram of a sample state showing direction of vehicle motion. Note that only the first 25 longitudinal grid cells are shown.

model in which these probabilities are 0.2 and 0.8, respectively. These transition functions are used to determine optimal merging policies within a variety of driving conditions.

3.4 Reward Function

The reward function $R(s, a)$ computes the expected reward for taking action a at state s. Two different reward function formulations are explored for this MDP to determine which reward structure better captures optimal driving behavior. The first formulation models rewards based on our prior knowledge of how we would expect autonomous vehicles to operate, directly encoding human values such as safety and mobility into this problem as a positive reward for merging, a penalty for merging close to the other vehicle, and a penalty for staying in the on-ramp. This reward function takes the form:

$$R_{prior\ knowledge}(s, a) = R_{merge} + R_{close} + R_{\neg merge}, \tag{1}$$

$$R_{merge} = \lambda_{merge} \ \ if \ \ a = 4 \tag{2}$$

$$R_{close} = -\lambda_{close}(\delta + |y_1 - y_e|)^{-1} \ \ if \ \ a = 4 \tag{3}$$

$$R_{\neg merge} = -1 \ \ if \ \ a \in \{1, 2, 3\} \tag{4}$$

where λ_{merge} and λ_{close} are parameters over which we can iterate to explore different merging behaviors, and $\delta = 0.1$ is a small buffer to ensure that R_{close} doesn't explode given overlapping y_1 and y_e values.

The second reward function formulation assumes no prior knowledge of human values and instead comprises a simple degree-one polynomial expression for the components of the state and the action as:

$$R_{polynomial}(s, a) = \alpha_1 x_e + \alpha_2 y_e + \alpha_3 y_1 + \alpha_4 a \tag{5}$$

where each coefficient α_i represents a parameter over which we can iterate to determine various optimal policies.

4 Solution Methods

Value iteration is utilized to iteratively compute optimal policies for this MDP given various transition and reward functions. The value at each state is computed as:

$$U_{k+1}(s) \leftarrow max_a[R(s, a) + \gamma \sum_{s'} T(s'|s, a)U_k(s')] \tag{6}$$

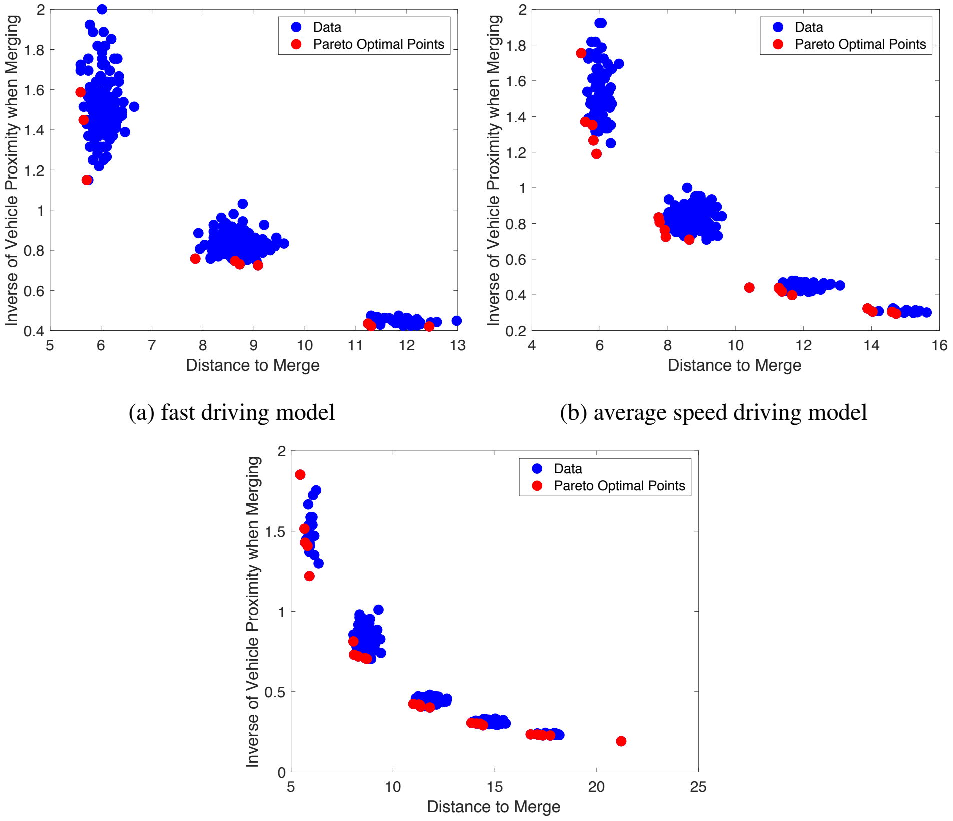
until the value function converges to $U^*(s)$ for each state. The optimal policy for a given state can then be extracted from the value function as:

$$\pi(s) = argmax_a[R(s,a) + \gamma \sum_{s'} T(s'|s,a)U^*(s')] \tag{7}$$

Each of these optimal policies is subsequently simulated in a stochastic environment in which the other vehicle moves longitudinally 1 step with a probability of 0.25, 2 steps with a probability of 0.5, and 3 steps with a probability of 0.25. The distance along the road when the vehicle merges and inverse of the longitudinal distance between the two vehicles when merging occurs is averaged for each policy over 100 simulations. This data point represents the competing objectives of mobility and safety; while we want the vehicle to merge as quickly as possible, we also want it to merge safely. This multi-objective optimization problem is formalized as minimizing the distance along the road at which point merging occurs and maximizing the gap between the two vehicles when merging, which is equivalent to minimizing the inverse of the gap when merging.

5 Results

The plots in Figure 2 show merging distance and inverse of vehicle proximity data points for each of the three transition dynamics for y_1 for the first reward model $R_{prior\ knowledge}(s,a)$. Each data point is computed for a $\{\lambda_{merge}, \lambda_{close}\}$ combination, having iterated over λ_{merge} values from 30 to 70 and λ_{close} values from 10 to 100. The clustering of data points in these plots can be explained by the discontinuous nature of the reward function. Given the conditional statements on the action space within $R_{prior\ knowledge}(s,a)$, only certain driving behaviors will be simulated over a large nested sweep of λ_i values. Within these plots, Pareto optimal points – values at which it is



(a) fast driving model

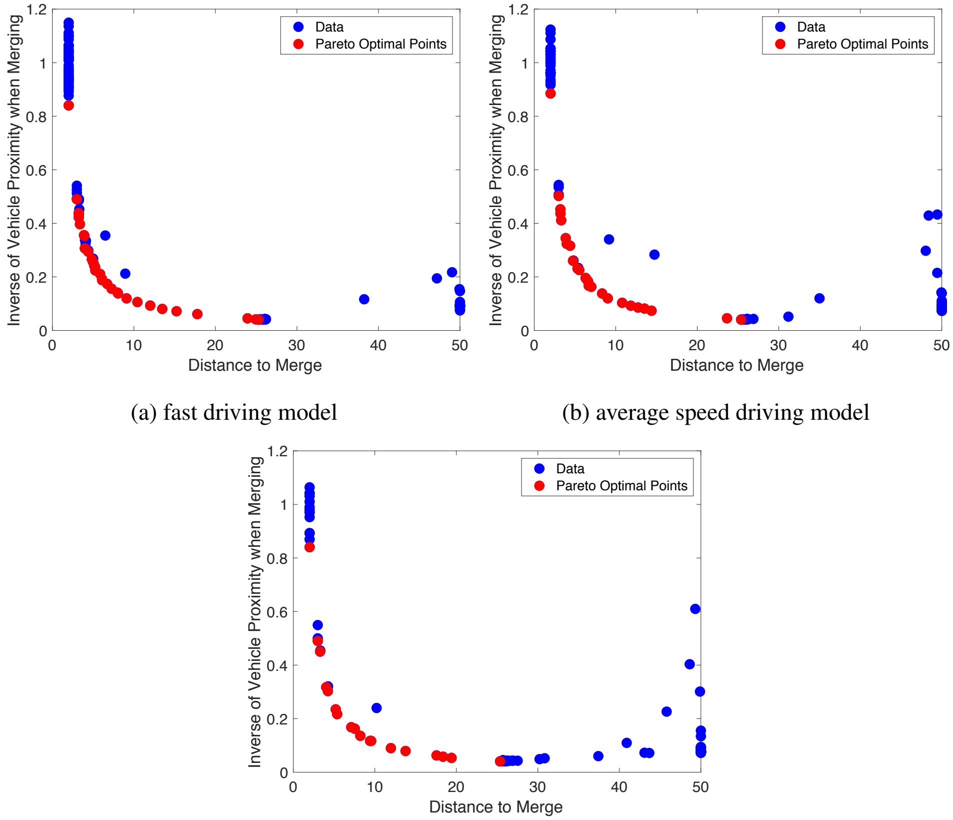
(b) average speed driving model

(c) slow driving model

Figure 2: Data points and Pareto optimal points for all three driving models based on prior knowledge reward function.

impossible to improve in the minimization of one objective without worsening our minimization of the other objective – have been colored red. One Pareto optimal point for plot (c) in Figure 2 which balances both objectives well with an x (mobility) value of 11 and y (safety) value of 0.42 occurs for $\lambda_{merge} = 48, \lambda_{close} = 60$. An animation of this optimal driving policy can be viewed here: `https://www.youtube.com/watch?v=icJz32L24JI`. We can trade off between safety and mobility by choosing λ_i values associated with the Pareto curve that connects these Pareto optimal points. Providing Pareto curves to vehicle policymakers and other important stakeholders can enable informed decisions on how best to program autonomous vehicles to meet our social expectations and ethical values.

In comparison, the plots in Figure 3 show the data points and Pareto optimal points for each of the three transition dynamics for y_1 for the second reward model $R_{polynomial}(s, a)$. To compute this data, we iterated over α_2, α_3, and α_4 values between -0.5 and 0.5, keeping α_1 at a value of -1. It is interesting to note that, compared to the results for the prior knowledge reward function, data for optimal policies computed with this reward function represents more regularly distributed points within the safety-mobility space. As a result of the continuous nature of this reward function, which provides rewards over the entire state and action space, a much more even spread of optimal driving policies – many of which are Pareto optimal – are computed. One Pareto optimal point for plot (c) in Figure 3 which balances both objectives well with an x (mobility) value of 8.2 and y (safety) value of 0.14 occurs for $\alpha_1 = -1, \alpha_2 = -0.17, \alpha_3 = 0.17, \alpha_4 = 0.17$. It is additionally worthwhile to note that, for both reward functions, the safety-mobility trade-off points shown in Figures 2 and 3 form a roughly Pareto optimal structure. This is particularly true for the second reward function $R_{polynomial}(s, a)$. By sweeping over different weights for both reward functions and computing optimal policies, these policies tend to follow a Pareto optimal shape when plotted as a function of



(a) fast driving model

(b) average speed driving model

(c) slow driving model

Figure 3: Data points and Pareto optimal points for all three driving models based on polynomial reward function.

merging distance and vehicle proximity given their optimality compared to other potential policies for a given set of weights.

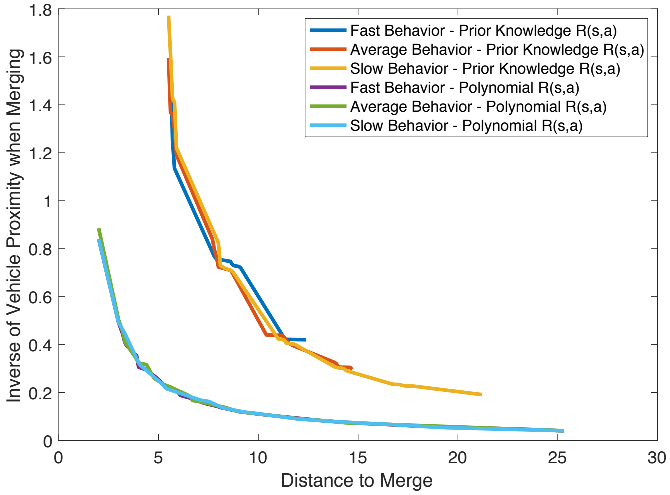


Figure 4: Pareto curves for both reward models and all three transition dynamics.

To compare among the various transition and reward functions, each set of Pareto optimal points is used to generate a Pareto curve as a simple linear interpolation. Figure 4 shows the Pareto curves for each transition-reward function pair. Among the Pareto curves for the prior knowledge formulation of the reward function, we see slight variations depending on the transition dynamics of the other vehicle used in computing an optimal merging policy. For slow behavior, the ego vehicle may traverse nearly half the grid space at $y_e = 22$ before merging in less risk averse optimal policies, while the ego vehicle may only travel longitudinally up to $y_e = 13$ before merging for fast behavior. For fast driving behavior, the ego vehicle merges a bit more quickly, but at the cost of safety when on a more aggressive portion of the Pareto curve. Further, the Pareto curves for the polynomial reward function far outperform those for the prior knowledge reward function. This is a very interesting result, as we'd expect a reward function incorporating more of our knowledge and intuition for this autonomous driving scenario to result in better policies. We can therefore conclude that there may exist better reward functions for capturing optimal driving policies than either the intuitive prior knowledge reward function or the polynomial reward function, which doesn't incorporate any human understanding of costs associated with safety and efficiency. A more rigorous approach to this problem using methods from IRL may enable the calculation of reward functions which result in policies that best optimize safety and mobility objectives.

6 Conclusion

We've formulated the problem of merging onto the highway as a discrete state and action space MDP and computed optimal policies via the value iteration algorithm. Faster driving models for the non-ego vehicle result in more aggressive merging policies, while slower driving models generate merging policies that slightly favor a longer time in the on-ramp before merging. Compared with the prior knowledge reward function, the polynomial reward function does much better at minimizing both safety and mobility objectives, even though it does not incorporate human knowledge of typical reward function structures for autonomous driving problems.

There are many future extensions of this problem that we'd like to explore. By incorporating both vehicles' velocity values into the state space, each vehicle's dynamics could be better represented, and other human values such as legality – i.e. minimizing the difference between the speed of each vehicle and the speed limit or a desirable highway merging speed – could be incorporated as an additional objective. Further, redefining this problem as a POMDP could enable a more interesting investigation into how other vehicles' partially observable position and speed and, not as directly observable, intent or inherent driving behavior affect optimal merging policies. Representing this problem with a continuous action and state space may result in a more refined simulation of specific dynamic maneuvers with tractable solutions for scenarios involving more than one non-ego vehicle. Lastly, reformulating this problem as an IRL task could facilitate a more thorough exploration of reward functions that best explain optimal driving behavior.

References

[1] S. Arora and P. Doshi, "A survey of inverse reinforcement learning: Challenges, methods and progress," *arXiv preprint arXiv:1806.06877*, 2018.

[2] A. Y. Ng, S. J. Russell, *et al.*, "Algorithms for inverse reinforcement learning.," in *Icml*, pp. 663–670, 2000.

[3] H. Bai, S. Cai, N. Ye, D. Hsu, and W. S. Lee, "Intention-aware online pomdp planning for autonomous driving in a crowd," in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pp. 454–460, IEEE, 2015.

[4] S. M. Thornton, F. E. Lewis, V. Zhang, M. J. Kochenderfer, and J. C. Gerdes, "Value sensitive design for autonomous vehicle motion planning," in *2018 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1157–1162, IEEE, 2018.

[5] H. Bai, D. Hsu, and W. S. Lee, "Integrated perception and planning in the continuous space: A pomdp approach," *The International Journal of Robotics Research*, vol. 33, no. 9, pp. 1288–1302, 2014.

[6] M. Bouton, A. Cosgun, and M. J. Kochenderfer, "Belief state planning for autonomously navigating urban intersections," in *Intelligent Vehicles Symposium (IV), 2017 IEEE*, pp. 825–830, IEEE, 2017.

[7] 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*, pp. 3004–3010, IEEE, 2017.

[8] J. Wei, J. M. Dolan, J. M. Snider, and B. Litkouhi, "A point-based mdp for robust single-lane autonomous driving behavior under uncertainties.," in *ICRA*, pp. 2586–2592, Citeseer, 2011.