

SAFETY AND EFFICIENCY IN AUTONOMOUS VEHICLES
THROUGH PLANNING WITH UNCERTAINTY

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF AERONAUTICS AND
ASTRONAUTICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Zachary Nolan Sunberg

August 2018

Abstract

To be effective, autonomous air and ground vehicles should maintain safety while accomplishing tasks efficiently in terms of time and other resources. Unfortunately, the objectives of safety and efficiency are fundamentally opposed because safety precautions prohibit some efficient actions. Moreover, the presence of uncertainty about the environment makes planning safe and efficient actions more difficult.

The Markov decision process (MDP) is a systematic framework for modelling sequential decision problems with outcome uncertainty, and the partially observable Markov decision process (POMDP) adds the additional ability to model state uncertainty. MDPs and POMDPs are suitable models for a wide range of situations that an autonomous vehicle might face. However, obtaining the exact solution to a general POMDP is an intractable problem. This thesis considers approximate MDP and POMDP solutions and seeks to quantify their utility for autonomous vehicles. Specifically, it contains four contributions.

First, the thesis analyzes the use of a certifiable safety constraint alongside approximate optimization in the context of unmanned aerial vehicle collision avoidance. In order to ensure safety, aerospace systems have particularly stringent certification requirements that likely preclude approximate randomized planning techniques capable of handling uncertainty. This work evaluates the performance price that comes with using a simple certified policy and shows that MDP and POMDP optimization can significantly reduce that price and improve both safety and efficiency simultaneously.

The second application chapter considers the effects of modeling uncertainty in a difficult lane changing task for a self-driving car. Specifically, the research estimates the value of planning with the internal states of other human drivers such as their

intentions and dispositions. While several other researchers have used internal-state-aware planning methods to interact with human drivers in desirable ways, they have not evaluated whether these methods offer a substantial quantitative improvement in performance over conventional approaches. This thesis shows that, in a simplified simulated setting, planning with internal states using a POMDP formulation can significantly improve both safety and efficiency simultaneously. Moreover, the thesis describes an experimental method for investigating other cases in which internal-state-aware planning may improve performance.

The benefits of POMDP planning can only be realized with algorithms that can handle real-world domains that are continuous and irregular. To that end, the third contribution of the thesis is a pair of new algorithms for solving POMDPs with continuous state, action, and observation spaces. These algorithms are motivated by analysis and numerical experiments that show that leading online POMDP solvers fail on continuous observation spaces because of the following two problems. First, the large observation space causes policy trees to become too wide and not deep enough. Second, the number of state particles used to represent beliefs collapses to one, causing overconfidence. The new algorithms, POMCPOW and PFT-DPW, handle these problems using progressive widening and weighted particle belief representations. Numerical experiments show that they are able to solve problems where previous methods fail.

The final contribution is a software package, POMDPs.jl, which uses the features of the Julia programming language to bring state-of-the-art POMDP solution methods to bear on problems defined through an interface that provides convenience and flexibility.

Acknowledgements

The biggest debt of gratitude that I owe to any single person for this thesis is to my advisor, Mykel Kochenderfer. In my opinion, the most admirable thing about Mykel is not the quality of work that he produces (though that is quite impressive), but his commitment and care for students as researchers and people. I will always think of his office as one of my favorite places of learning on campus. He took me on as a student during a time of struggling in my research and helped me finish strong.

I also owe a great deal to Marco Pavone, who was my advisor for the first half of this endeavor. Additionally, my thanks goes out to Dan DeBra, who brought me here to Stanford, and Suman Chakravorty and Jon Rogers who were instrumental in my early growth as a researcher at Texas A&M. I am also thankful to the other members of my defense committee: Mac Schwager, Chris Gerdes, and Dorsa Sadigh.

A PhD is not a disembodied pursuit of knowledge, but a journey undertaken by a human with a heart and soul and flaws and weaknesses. As such, I must thank my community here at Stanford for their care of me as a human exploring God's creation. Grace Presbyterian church provided a place of connection and growth for me, and my friends gave me constant support. In particular, two friends, Jeff Reid and Daniel Galbraith, should be noted for their faithfulness over the entire time I was here. Moreover, I certainly would not be here without my parents and sister and the rest of my family and friends who have been shaping me for my entire life.

I was fortunate to be a part of two outstanding labs during my time here, the Autonomous Systems Lab and the Stanford Intelligence Systems Lab, both of which gave me comraderie and sharpened my technical skills. I'll remember both the hard work and fun that I had in these places.

The work in this thesis would not have been possible without the generous financial support that I received from a variety of sources including MIT Lincoln Laboratory, the National Science Foundation, the U.S. Naval Research Laboratory, and Toyota Research Institute. I must also recognize the leadership and staff of Stanford University and the Aeronautics and Astronautics department. They have created a remarkable environment for learning.

Contents

Abstract	iv
Acknowledgements	vi
1 Introduction	1
1.1 Autonomous Vehicles	1
1.1.1 Benefits of Autonomous Transportation	1
1.1.2 Current Progress	3
1.1.3 Remaining Challenges	3
1.2 Decision Making Under Uncertainty	5
1.2.1 Optimization objective	5
1.2.2 Uncertainty in Decision Making	7
1.2.3 Markov Decision Processes	9
1.2.4 Partially Observable Markov Decision Processes	10
1.2.5 Value Iteration	11
1.2.6 Monte Carlo Tree Search	12
1.2.7 Particle Filtering	14
1.2.8 Approximate Solutions to POMDPs	15
1.2.9 QMDP	17
1.3 Contributions and Outline	18
2 Trusted and Optimized UAV Collision Avoidance	20
2.1 Collision Avoidance for UAVs	20
2.2 MDP Models	23

2.2.1	Model Assumptions	24
2.2.2	Vehicle States and Dynamics	25
2.2.3	Transition Function	26
2.2.4	Reference Trusted Resolution Logic	27
2.2.5	Action Spaces and Control Systems	28
2.2.6	Reward	30
2.2.7	Problem statement	31
2.3	Solution Approach	31
2.3.1	Approximate Value Iteration	32
2.3.2	Post Decision Value Function Extraction	34
2.3.3	Online Policy Evaluation	36
2.3.4	Selection of Features	36
2.4	Results	40
2.4.1	Policies	40
2.4.2	Numerical Performance Evaluation	42
2.5	Discussion	46
3	Planning with Internal States in Driving	48
3.1	Human Interaction in Autonomous Driving	48
3.2	Freeway Driving POMDP	50
3.2.1	Driver Modeling	52
3.2.2	Physical Dynamics	53
3.2.3	Action Space for Crash-Free Driving	55
3.2.4	Reward Function and Objectives	56
3.2.5	Initial Scenes	57
3.3	Solution Approaches	57
3.3.1	Approach 1: Assume normal behavior	57
3.3.2	Approach 2: Model all uncertainty as outcome uncertainty (Naive MDP)	57
3.3.3	Approach 3: Mean Model Predictive Control	58
3.3.4	Approach 4: QMDP	59

3.3.5	Approach 5: POMCPOW	59
3.4	Results	59
3.4.1	Driver Model Distribution Scenarios	60
3.4.2	Pareto Front Comparison	63
3.4.3	Correlation comparison	66
3.4.4	Robustness	67
3.5	Discussion	70
4	Online Algorithms for Continuous POMDPs	73
4.1	Background	74
4.2	Algorithms	75
4.2.1	POMCP-DPW	76
4.2.2	PFT-DPW	79
4.2.3	POMCPOW	81
4.2.4	Discretization	81
4.2.5	Observation Distribution Requirement	83
4.3	Experiments	83
4.3.1	Laser Tag	84
4.3.2	Light Dark	84
4.3.3	Sub Hunt	86
4.3.4	Van Der Pol Tag	88
4.3.5	Multilane	89
4.3.6	Discretization granularity	89
4.3.7	Hyperparameters	90
4.4	Discussion	91
5	A Julia Framework for POMDPs	93
5.1	Challenges for POMDP-solving software	94
5.1.1	Speed	94
5.1.2	Flexibility	94
5.1.3	Ease of Use	95
5.2	Previous frameworks	96

5.3	Architecture	97
5.3.1	Concepts	97
5.3.2	Interfaces	98
5.4	Examples	100
5.4.1	Problem	100
5.4.2	Solver	103
5.4.3	Simulator	103
5.5	Discussion	104
6	Summary and Future Work	106
6.1	Contributions and Summary	106
6.2	Future Work	108
A	Proof of Theorem 1	111

List of Tables

2.1	Parameters for numerical experiments	45
3.1	IDM and MOBIL parameters for different driver types.	61
3.2	Various simulation parameters	62
4.1	Experimental results	85
4.2	Hyperparameters used in experiments	91

List of Figures

2.1	Trusted resolution logic	22
2.2	Intruder interpolation grid	38
2.3	Goal interpolation grid	39
2.4	UAV Collision avoidance policy visualizations	41
2.5	Intruder initial conditions	43
2.6	Policy performance comparison	44
3.1	Lane changing scenario	50
3.2	Lane changing action space	55
3.3	Performance curves at correlation extremes	64
3.4	Performance curves at correlation 0.75	65
3.5	Average hard braking frequency and success rate	67
3.6	Performance variation with Θ correlation	68
3.7	Parameter correlation robustness study	69
3.8	Parameter domain robustness study	70
4.1	POMCP tree on a continuous observation space	74
4.2	POMCP-DPW and POMCPOW tree structure comparison	78
4.3	Laser Tag benchmark	84
4.4	Light Dark problem	86
4.5	Sub Hunt problem	87
4.6	Van Der Pol tag problem	88
4.7	Discretization granularity studies	90

5.1	POMDPs.jl concepts	98
5.2	POMDPs.jl interfaces	100
5.3	POMDPs.jl requirements example	101

Chapter 1

Introduction

Safety and efficiency are two important goals for technology in many domains. This thesis focuses on the autonomous vehicle domain, an application area that is currently growing quickly. Before outlining specific research results, this chapter gives an introduction to some benefits of and challenges for autonomous vehicles and some of the basic theoretical structures for approaching the problem of controlling them.

1.1 Autonomous Vehicles

Until now, vehicles have been almost exclusively operated by humans. However, recent advances in sensing technology, computing power, mapping, data processing, and connectivity have made it possible to begin developing vehicles that are *autonomous*, that is, they can perform a transportation task without requiring a human to control them.

1.1.1 Benefits of Autonomous Transportation

Autonomous vehicles have the potential to bring a wide range of benefits. The first of these benefits is **improved safety**. Around the world, more than 1.2 million people are killed every year by vehicles [37], and, according to a study concluded in 2007, the critical reason for more than 90% of motor vehicle crashes in the United

States is driver error [94]. Since they are not affected by many of the crash-causing deficiencies that human drivers have, including distraction, tiredness, and impairment due to drugs or alcohol, autonomous vehicles have the potential to greatly reduce the number of injuries and deaths.

Another potential benefit of autonomous vehicles is **increased collective efficiency**. Currently, middle- and low-income nations lose up to 3% of GDP due to road vehicle accidents, many of which could be mitigated with autonomy [37]. Moreover, autonomy has the potential to enable new models of vehicle usage for cities. For example, a city-wide mobility-on-demand (MOD) system consisting of purpose-built electric autonomous vehicles could provide a much more efficient system for commuting than the current fleet of individually-owned gas-powered vehicles optimized for factors other than commuting efficiency. The barriers to adopting such a system may be greatly reduced by autonomous vehicles' ability to, for instance, automatically travel to a charging station or rebalance to places where additional vehicles are needed. A recent study concluded that Manhattan taxi demand could be met with a fleet of autonomous cars 60% the size of the current taxi fleet [112].

The third benefit is the potential for **reduced personal transportation costs** in terms of time and stress. In the United States, workers commute an average of more than 25 min in each direction [68], and more than three quarters drive alone [67]. Thus a large fraction of American workers use an hour of their day focused on operating a vehicle. With autonomous vehicles, this stressful time could be eliminated and replaced with productive time or rest.

The final potential benefit that will be listed here is **expanding access to mobility**. The requirement for a capable vehicle operator prevents many people from using cars. For example, many people with permanent disabilities such as blindness or seizure disorders or even temporary injuries are precluded from the flexible personal point-to-point transportation afforded by cars, and this increases their dependence on caretakers. Autonomous cars would give these people safe and flexible personal transportation.

1.1.2 Current Progress

There has been a great deal of progress towards autonomy. For example, several companies including Amazon [91] and Alphabet [86], are pursuing delivery systems consisting of autonomous aerial drones. A focal point of early ground vehicle development was the 2007 DARPA Urban Challenge [19], which was the first major demonstration of self-driving cars operating in a simulated urban environment. Since that time, significant progress has been made and testing operations in a small number of urban environments has become routine [26].

1.1.3 Remaining Challenges

Despite the considerable progress made towards autonomy, there are currently no large-scale commercial applications of autonomous vehicles, and several challenges remain. This section does not give an exhaustive list of the challenges remaining for deployment of autonomous vehicles, but it outlines a few and discusses how this thesis relates to them.

Technical

Some of the largest remaining technical challenges for autonomous vehicle development are challenges of generalization. Many of the components for navigating within a closed test space had already been developed by the time of the DARPA urban challenge in 2007 [19]. An industry leader has noted that “it’s relatively easy to master the first 90 % of driving where you’re traveling on freeways, navigating light city-street traffic, or making your way through simple intersections” [26]. However, autonomous vehicle testing is still limited to relatively small well-mapped environments because it is difficult to generalize solutions to new domains.

In order to handle a larger and more diverse set of circumstances, autonomous vehicle designers must both gather more data [26] and design control systems to operate in the new environment. One approach to control system design is to iterate with a two step process consisting of (1) specifying a control policy by hand, and (2)

evaluating these actions against some criteria. Examples of this approach are hand-tuned linear control systems and the specified behaviors of some of the urban challenge competitors (e.g. Urmson, Anhalt, Bagnell, *et al.* [107]). A second approach is to use algorithms that systematically and automatically determine a good control policy according to some criterion. While the first approach is adequate for some simple constrained domains, the second is more likely to work in regimes that the designers have not explicitly tested. Moreover, the second approach reduces the number of tasks required to adapt to a new operational environment from two (identifying a model and designing a control system) to one (identifying a model). The algorithms and applications in this thesis belong to the latter family and improve our ability to apply the approach to problems with complex and uncertain dynamics.

There are many other technical challenges including perception in adverse conditions, testing, security, cost reduction, linking perception with reasoning, and others, but they are not directly addressed in this work.

Ethical, Legal, and Economic

In addition to the technical challenges, there are a wide range of ethical, legal, and economic challenges, including liability assignment, job displacement, and discrimination against certain populations, that have the potential to hinder the success of autonomous vehicles [20]. This thesis does not seek to systematically address moral or legal challenges, but it does speak to one in particular. Autonomous vehicles must often make decisions that balance the interests of multiple people. For instance, in addition to the performance criteria that the vehicle operator desires, there may be *externalities* that affect the well-being of other people. There is often a tradeoff between these objectives. For instance, traveling with increased speed will allow the owner to transport passengers or cargo faster, but it may endanger bystanders. The results in this thesis show that, for the specific competing goals of safety and efficiency, both can be simultaneously improved by using better models and algorithms (see Chapters 2 and 3). It is likely that this principle will also apply to other trade-offs. Better algorithmic and modeling approaches may be an answer in cases where competing ethical, legal, and economic principles are at odds.

1.2 Decision Making Under Uncertainty

The task of controlling an autonomous vehicle is a task of decision making under uncertainty. That is, the vehicle control system must decide what actions to take to achieve good performance, even when there is uncertainty about the world around it and what will happen in the future. There are two ingredients needed to formulate such a problem. The first is a criterion for good performance, and the second is a model of what is known about the world and how it will evolve.

1.2.1 Optimization objective

This thesis casts decision making problems as optimization problems. That is, the criterion for good performance is that the actions chosen maximize an objective function,¹

$$\underset{\pi}{\text{maximize}} \quad J(\pi) \tag{1.1}$$

where π determines what actions are taken depending on the inputs. The concrete structures for π and J are presented in Section 1.2.3, but a brief introduction is given here.

To define an objective function, an engineer must map the desires of human passengers or abstract societal goals like those described in Section 1.1.1 into a well-defined mathematical function. This is sometimes straightforward for individual goals such as minimizing the time to reach a destination. However, combining multiple goals is often a challenge because an optimization problem is not, in general, well-posed if it has multiple objective functions.

One solution to this problem is to optimize a single objective, J_0 , and constrain

¹This thesis makes no distinction between objective *functionals* that map from spaces of functions to scalar values and objective functions that map from scalar or vector spaces to scalar values. The term “objective function” will be used for both.

all other objectives, J_i , to meet specified thresholds, D_i [2],

$$\begin{aligned}
 & \underset{\pi}{\text{maximize}} && J_0(\pi) \\
 & \text{subject to} && J_1(\pi) \geq D_1 \\
 & && J_2(\pi) \geq D_2 \\
 & && \dots \\
 & && J_n(\pi) \geq D_n.
 \end{aligned} \tag{1.2}$$

However many off-the-shelf decision-making algorithms must be modified to handle this formulation, and choosing the thresholds adds another step to the development process.

Another alternative is to create a single objective function that is a weighted sum of the functions corresponding to the individual rewards,

$$\underset{\pi}{\text{maximize}} \quad J_0(\pi) + \lambda_1 J_1(\pi) + \lambda_2 J_2(\pi) + \dots + \lambda_n J_n(\pi), \tag{1.3}$$

where λ_i are the relative weights. This scalarization approach allows conventional decision-making algorithms to be used, but choosing appropriate weights before solving the problem is often difficult.

Another challenge is that human preferences can be difficult to quantify and encode into a reward function. For these cases, inverse reinforcement learning [64, 85] or preference elicitation [63] can be used to determine a suitable reward function based on data recording how humans act.

This thesis does not make significant contributions to understanding how to choose appropriate objectives, but instead adopts the weighted sum scalarization approach to combine a few important objectives. In particular, the investigation focuses on the goals of safety and efficiency, which are balanced using a scaling factor, λ ,

$$\underset{\pi}{\text{maximize}} \quad J_E(\pi) + \lambda J_S(\pi). \tag{1.4}$$

It adopts a stance that is agnostic to the value λ by focusing on Pareto frontier

analysis.

A solution is said to be *Pareto-optimal* if no objective can be improved without adversely affecting another objective. In particular, every solution to problem (1.4) for a positive value of λ is Pareto-optimal, though there may also be additional Pareto-optimal points that do not correspond to solutions of this problem. The *Pareto front* is the set of all Pareto-optimal solutions.

Since exact solutions to the problems studied in this thesis are intractable, it is impossible to reliably generate points on the true Pareto front. Instead, approximate Pareto fronts are constructed by connecting approximate solutions to (1.4) with straight lines (see Fig. 2.6 for an example), and conclusions about different algorithms can be reached by comparing the resulting curves generated by different algorithms. Given two algorithms, A and B, if the Pareto front for A is in a better position than that of B for most or all values of λ , then it is possible to argue that algorithm A is superior to algorithm B without committing to a particular value of λ .

1.2.2 Uncertainty in Decision Making

The second primary ingredient to a decision making problem is a model of the randomness in the mapping from actions to the objective function. This thesis is concerned with three types of uncertainty: outcome, model, and state. Although the definitions of these types are not mathematically formal, they are useful for conceptual understanding. The distinctions between the types are defined based on the concept of “Markov state”.

Markov state

Consider the case where the configuration of the world changes over time according to a dynamics model. This thesis only considers models where these changes occur at specific discrete time intervals. A *Markov state*, or simply *state*, consists of the world configuration and possibly some other information specially chosen so that the state at the next time step depends only on the current state and not on any previous states. This property of independence from past history is known as the *Markov*

property. In general, the state may be a random variable, and, in this case, the next state must be stochastically independent of all previous states when conditioned on the current state.

Outcome Uncertainty

The simplest type of uncertainty is outcome uncertainty. Outcome uncertainty is uncertainty in the future state even when the current state and model are known exactly. In other words, it refers to the inherent stochasticity in the model. The distinguishing property of outcome uncertainty is that nothing can be learned by the agent to reduce it.

Model Uncertainty

A second category for uncertainty is model uncertainty. In some cases, the dynamic model of the system is not known to the agent, and this can be a source of uncertainty in the future of the state that compounds the outcome uncertainty. The distinction is that, over time the agent can observe how the state changes to learn more about the model and reduce this type of uncertainty. Methods that attempt to optimize an objective in the face of this type of uncertainty are usually referred to as *reinforcement learning* methods [101].

State Uncertainty

A third type of uncertainty is state uncertainty. When there is state uncertainty, the agent cannot directly observe the state, and instead receives only observations that offer clues about it. However, it may reason about what the state could be and make better decisions based on this partial knowledge. The distinction between state and model uncertainty is that the state may change over time, while the model does not. Model uncertainty can be represented as state uncertainty by augmenting the state to include the unknown model parameters.

1.2.3 Markov Decision Processes

The Markov decision process (MDP) is a mathematical formalism that can represent a wide range of sequential decision making problems [10, 52]. In an MDP, an agent takes *actions* that affect the state of the system and collects *rewards* based on the states and actions. The MDP formalism is able to represent outcome uncertainty with a stochastic state transition model. Typically the state at time t will be denoted s_t , and the action at time t denoted a_t , however in some contexts where only one step of time is relevant, the subscripts will be dropped for the sake of clarity and brevity, and s' will be used to denote the next state that the system transitions into.

Formally, an MDP is defined by the 5-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma)$. The state space, \mathcal{S} , is the set of all possible states. The action space, \mathcal{A} , is the set of all actions available to the agent. The transition model, \mathcal{T} , represents the likelihood of different transitions, where $\mathcal{T}(s' | s, a)$ denotes the probability that the system will transition to state s' given that action a is taken in state s . The reward function, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$ represents the rewards received while interacting in the environment, where $\mathcal{R}(s, a, s')$ denotes the reward for transitioning from s to s' when action a is taken. Finally, γ governs how reward is discounted in the future.

The objective in an MDP is to find a policy, $\pi : \mathcal{S} \rightarrow \mathcal{A}$, that maps each encountered state to an action and, when $a_t = \pi(s_t)$, maximizes the objective function. Because of the outcome uncertainty, the rewards that a policy will achieve are random variables. Thus, for the optimization problem to be well-posed, the objective must be a function of the distribution of the rewards. For a Markov decision process, the objective is to maximize the cumulative expected reward,²

$$J(\pi) = E \left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, \pi(s_t), s_{t+1}) \right]. \quad (1.5)$$

It is important to note that it is sufficient to consider only the class of deterministic Markov policies, that is, policies that map each state to a single action, when searching

²Sequential decision problems where the objective involves maximizing a measure other than the expectation of the reward have been studied, but are much less common, and their solution is sometimes much more difficult than in the expectation case [22].

for an optimal solution. It has been proven that, for a policy of any structure, e.g. a policy that is stochastic or a policy that depends on more than just the state, there exists a Markov policy that can attain the same objective value. Consequently, it is always possible to find a deterministic Markov policy that achieves the optimal objective value [2].

The objective of an MDP is sometimes discussed in terms of the optimal value function $V^* : \mathcal{S} \rightarrow \mathbb{R}$, which is the cumulative expected reward given that the system starts in a specified state. The optimal state action value function, $Q^*(s, a)$, is defined as the expectation of the future cumulative reward given that the agent starts in state s , immediately takes action a , and then follows the optimal policy [10, 52, 2].

1.2.4 Partially Observable Markov Decision Processes

A partially observable Markov decision process (POMDP) is similar to an MDP except that the agent cannot directly observe the state. Instead, the agent only has access to observations that are generated probabilistically based on the actions and latent true states. In this way, a POMDP is able to represent state uncertainty in addition to the outcome uncertainty that can be encoded in an MDP. A POMDP is defined by the 7-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \mathcal{O}, \mathcal{Z}, \gamma)$, where \mathcal{S} , \mathcal{A} , \mathcal{T} , \mathcal{R} , and γ have the same meaning as in an MDP. Additionally, \mathcal{O} is the observation space, and \mathcal{Z} is the observation model. $\mathcal{Z}(o | s, a, s')$ is the probability or probability density of receiving observation o in state s' given that the previous state and action were s and a [10, 52, 47].

Information about the state may be inferred from the entire history of previous actions and observations and the initial information, b_0 . Thus, in a POMDP, the agent's policy is a function mapping each possible history, $h_t = (b_0, a_0, o_1, a_1, o_2, \dots, a_{t-1}, o_t)$ to an action. In some cases, each state's probability can be calculated based on the history. This distribution is known as a *belief*, with $b_t(s)$ denoting the probability of state s .

The belief is a sufficient statistic for optimal decision making. That is, there exists a policy, π^* such that, when $a_t = \pi^*(b_t)$, the expected cumulative reward or “value function” is maximized for the POMDP [47, 52]. Given the POMDP model, each

subsequent belief can be calculated using Bayes' rule according to

$$b'(s') = \frac{\int_{s \in \mathcal{S}} \mathcal{Z}(o | s, a, s') \mathcal{T}(s' | s, a) b(s) ds}{\int_{s' \in \mathcal{S}} \int_{s \in \mathcal{S}} \mathcal{Z}(o | s, a, s') \mathcal{T}(s' | s, a) b(s) ds ds'}. \quad (1.6)$$

When the state space is discrete, the integrals may be replaced with sums.

Unfortunately, it has been shown that even the class of finite-horizon POMDPs is PSPACE-complete, indicating that it is extremely unlikely that efficient general exact algorithms for large problems will be discovered [76]. Because of this, approximate algorithms are often used.

Generative Models

For many problems, it can be difficult to explicitly determine or represent the probability distributions \mathcal{T} or \mathcal{Z} . Some solution approaches, however, only require samples from the state transitions and observations. A generative model, G , stochastically generates a new state, reward, and observation in the partially observable case, given the current state and action, that is $s', r = G(s, a)$ for an MDP, or $s', o, r = G(s, a)$ for a POMDP. A generative model implicitly defines \mathcal{T} and \mathcal{Z} , even when they cannot be explicitly represented.

Belief MDP

Every POMDP is equivalent to an MDP where the state space of the MDP is the space of possible beliefs. The reward function of this "belief MDP" is the expectation of the state-action reward function with respect to the belief. The Bayesian update of the belief serves as a generative model for the belief space MDP.

1.2.5 Value Iteration

Value iteration is the simplest offline method for solving MDPs. It relies on the Bellman equation, which characterizes the optimal value function,

$$V^*(s) = \max_{a \in \mathcal{A}} \{ \mathcal{R}(s, a) + \mathbb{E} [V^*(s') | s, a] \}. \quad (1.7)$$

For discrete problems, the unique solution to the Bellman equation, V^* , can be found through repeated application of the Bellman operator,

$$\mathcal{B}[V](s) = \max_{a \in \mathcal{A}} \{ \mathcal{R}(s, a) + \mathbb{E}[V(s') \mid s, a] \}. \quad (1.8)$$

When \mathcal{B} is applied iteratively, i.e. $V_{k+1} = \mathcal{B}[V_k]$, the value function will eventually reach a fixed point when $V_{k+1} = V_k$.³ This point is the optimal value function, and the process of finding the value function in this manner is known as value iteration [10].

Once the optimal value function has been found, the optimal policy may be easily extracted according to

$$\pi^*(s) = \operatorname{argmax}_{a \in \mathcal{A}} \{ \mathcal{R}(s, a) + \mathbb{E}[V^*(s') \mid s, a] \}. \quad (1.9)$$

1.2.6 Monte Carlo Tree Search

In many cases, the state and action spaces of a problem are so large that computing a complete solution offline is intractable. Instead, approximate solutions valid only for the current state may be calculated online during execution in the environment. One way of accomplishing this is to create a policy tree consisting of alternating layers of state nodes and action nodes that evaluates every possible future trajectory. Unfortunately, it is typically infeasible to construct a sufficiently deep tree because construction is exponentially complex with respect to the depth [52].

Monte Carlo tree search (MCTS) is a method that uses Monte Carlo simulations to incrementally construct only important parts of the policy tree. By estimating the state-action value function, $Q(s, a)$, for each of the action nodes within the tree and focusing further exploration on nodes with high values, it can achieve good results without attempting the exponentially complex task of constructing the entire policy tree. MCTS is widely studied and has proven to be effective in a variety of contexts [17]. Only a generative model, G , is required by the algorithm. Although there are many variations of this algorithm, they are built around a common Monte

³The uniqueness of the fixed point and convergence to that fixed point are guaranteed because the Bellman operator is a contraction mapping [46].

Carlo simulation process that is repeated n times. Each simulation consists of the following four steps:

1. *Search.* In the initial phase of the simulation, the policy defined by the tree is used. At each state node, a selection criterion based on Q is used to choose a favorable action, and the tree is traversed through the node and to the next state node determined by G .
2. *Expansion.* Eventually, the simulation reaches an action node that does not have any children. At this point, a new state is sampled with G and a corresponding node created along with children corresponding to each action.
3. *Rollout.* After the expansion step, the simulation is continued with a rollout policy, often consisting of randomly selected actions, until the future accumulated reward will be negligible because of the compounding discount factor.
4. *Update.* Once the simulation has terminated, the estimates of $Q(s, a)$ at each of the visited action nodes are updated with the discounted reward received during the simulation after visiting the node.

This approach builds a tree asymmetrically favoring regions of the state and action spaces that are likely to be visited when the optimal policy is executed.

Upper Confidence Trees

The selection criterion used to choose actions in the search phase is very important. It must balance favoring actions with large Q values that are expected to yield good results with exploration of new actions. A widely used approach for this is the upper confidence bound for trees (UCT) algorithm [17]. At each state node, it chooses the action that maximizes the upper confidence bound

$$UCB(s, a) = Q(s, a) + c \sqrt{\frac{\log N(s)}{N(s, a)}} \quad (1.10)$$

where $N(s, a)$ is the number of times the action node has been visited, $N(s) = \sum_{a \in \mathcal{A}} N(s, a)$, and c is a problem-specific parameter that governs the amount of

exploration in the tree. The second term causes the algorithm to favor actions that have been taken less often.

Double Progressive Widening

In cases where the action and state spaces are large or continuous, the MCTS algorithm will produce trees that are very shallow. In fact, if the action space is continuous, the UCT algorithm will never try the same action twice (it is understood that, if $N(s, a) = 0$ then $UCB(s, a)$ in (1.10) is infinite, so untried actions are always favored). Moreover, if the state space is continuous and the transition probability density is finite, the probability of sampling the same state twice from G is zero. Because of this, simulations will never pass through the same state node twice and a tree below the first layer of state nodes will never be constructed.

In progressive widening, the number of children of a node is artificially limited to kN^α , where N is the number of times the node has been visited and k and α are parameters chosen for the problem [23]. Originally, progressive widening was applied to the action space, and was found to be especially effective when a set of preferred actions was tried first [17]. The term *double* progressive widening refers to progressive widening in both the state and action space. When the number of state nodes is greater than the limit, instead of simulating a new state transition, one of the previously generated states is chosen with probability proportional to the number of times it has been previously generated.

1.2.7 Particle Filtering

Aside from a few special cases, for example when the system dynamics are linear and the transition and observation distributions are Gaussian, the integrals in the Bayesian belief update (1.6) are impossible or difficult to solve analytically. Thus, numerical approaches must be used. A popular technique for this is particle filtering, usually incorporating domain-specific heuristic modifications to prevent problems such as particle depletion [104].

Sequential importance resampling, one of the most common and effective variations, requires a state generative model, G_s that can sample from the transition distribution and an explicit representation of the observation distribution, $\mathcal{Z}(\cdot | s, a, s')$. The belief is approximated with a set of m particles, $\{s_i\}_{i=1}^m$ and associated weights, $\{w_i\}_{i=1}^m$. The probability of each state is approximated by the sum of the weights corresponding to particles with that state value,

$$b(s) \approx \sum_{i=1}^m w_i \delta_s(s_i) \quad (1.11)$$

where $\delta_s(\cdot)$ is the Dirac delta function centered at s . A belief update is approximated by sampling m states $\{\tilde{s}_i\}_{i=1}^m$ from the collection of particles with probability proportional to the associated weight, generating a particle, $\tilde{s}'_i = G(\tilde{s}_i, a)$ for each of these states, and finally setting the new weight proportional to the probability of generating the measured observation with this state, $w'_i \propto \mathcal{Z}(o | \tilde{s}_i, a, \tilde{s}'_i)$.

1.2.8 Approximate Solutions to POMDPs

Considerable progress has been made in solving large POMDPs. Initially, exact offline solutions to problems with only a few discrete states, actions, and observations were sought by using value iteration and taking advantage of the convexity of the value function [47], although solutions to larger problems were also explored using Monte Carlo simulation and interpolation between belief states [103]. Many effective offline planners for discrete problems use point based value iteration, where a selection of points in the belief space are used for value function approximation [58]. Offline solutions for problems with continuous state and observation spaces have also been proposed [8, 16].

There are also various solution approaches that are applicable to specific classes of POMDPs, including continuous problems. For example, Platt, Tedrake, Kaelbling, and Lozano-Perez [78] simplify planning in large domains by assuming that the most likely observation will always be received, which can provide an acceptable approximation in some problems with unimodal observation distributions. Moreover,

Marchant, and Ramos [71] solve a monitoring problem with continuous spaces with a Gaussian process belief update. Hoey and Poupart [43] propose a method for partitioning large observation spaces without information loss, but demonstrate the method only on small state and action spaces that have a modest number of conditional plans. Other methods involve motion-planning techniques [69, 79, 18]. In particular, Agha-Mohammadi, Chakravorty, and Amato [1] present a method to take advantage of the existence of a stabilizing controller in belief space planning. Van Den Berg, Patil, and Alterovitz [108] perform local optimization with respect to uncertainty on a pre-computed path, and Indelman, Carlone, and Dellaert [45] devise a hierarchical approach that handles uncertainty in both the robot’s state and the surrounding environment.

General purpose online algorithms for POMDPs have also been proposed. These algorithms are mostly derivatives of the MCTS algorithm for MDPs (Section 1.2.6). A conceptually straightforward way to solve a POMDP using MCTS is to apply it to the corresponding belief MDP. Indeed, many tree search techniques have been applied to POMDP problems in this way [81]. However, when the Bayesian belief update is used, this approach is computationally expensive.

A major improvement in efficiency came with the introduction of the partially observable Monte Carlo planning (POMCP) algorithm [92]. It is based on UCT, but can tackle problems many times larger than its predecessors because it uses state trajectory simulations, rather than full belief trajectories, to build the tree. Each node in the POMCP tree corresponds to an action- or observation-terminated history. At each iteration, a state is sampled from the current belief and simulated through the nodes that match the history as it is generated. Thus, the belief at each observation-terminated history node is effectively represented by the collection of states for the appropriate time step in the trajectories that match the history. This is similar to using a rejection sampling unweighted particle filter. It should be noted that when this thesis uses the term POMCP, it refers only to the UCT-based decision-making algorithm called PO-UCT by Silver, Huang, Maddison, *et al.* [92] and not to the belief update scheme that re-uses simulations from the planning stage.

Determinized sparse partially observable tree (DESPOT) is a similar approach

that attempts to achieve better performance by analyzing only a small number of random outcomes in the tree [97]. Additionally, an algorithm called adaptive belief tree (ABT) was designed specifically to accommodate changes in the environment without having to replan from scratch [60].

Some very recent methods, e.g. QMDP-Net [49], have attempted to solve POMDPs by training recurrent neural networks.

1.2.9 QMDP

One POMDP approximation that is particularly useful in many contexts, but known to be sub-optimal for some, is the QMDP approximation. This approximation is based on an algorithm proposed by Littman, Cassandra, and Kaelbling [65]. However, in this thesis, it will refer to a broader class of approximations. Specifically, it refers to any policy where the action maximizes the expected Q_{MDP} value for the belief.

Let Q_{MDP} be the optimal state-action value function for the fully observable MDP at the core of the POMDP. Then the expected Q_{MDP} value for a belief action pair is $Q_{\text{MDP}}(b, a) = \int_{s \in \mathcal{S}} Q_{\text{MDP}}(s, a) b(s) ds$. Since it chooses the action that maximizes this value, it is easy to see that the QMDP approximation is the optimal solution to a problem with partial observability on the current step, but that subsequently becomes fully observable.

Because the QMDP approximation assumes that all information about the state of the problem will become available on the next step, it has no incentive for learning about the state, and hence cannot choose to take costly information-gathering actions, even if they are part of the optimal solution to the POMDP. Nevertheless, the approximation is useful in many domains because it is much easier to compute than the full POMDP solution — it requires only the solution to the fully observable MDP, which can usually be calculated efficiently using value iteration or another method. Moreover the QMDP solution has acceptable quality for many domains (see, for example, Table 4.1).

1.3 Contributions and Outline

The body of this thesis is laid out in four chapters. Each contains one or more contributions.

Chapter 2 analyzes the use of certifiable trusted resolution logic (TRL) alongside approximate optimization in the context of unmanned aerial vehicle collision avoidance. Specifically, the price of certifiability is quantified by observing the gap between the Pareto fronts for two approaches: (1) using the TRL with varying levels of conservativeness and (2) directly solving the MDP representation of the problem. Simulation results then demonstrate that this gap can be largely closed without sacrificing certifiability by using the TRL as a safety constraint on the action space and solving the resulting MDP.

Chapter 3 considers the effects of modeling uncertainty in a difficult lane changing task for a self-driving car. Specifically, simulations are used to estimate the Pareto fronts for planning with several approximations of the internal states of other human drivers such as their intentions and dispositions. While several other researchers have used internal-state-aware planning methods to interact with human drivers in desirable ways, they have not evaluated whether these methods offer a substantial quantitative improvement over conventional approaches. In a simplified simulated setting, planning with internal states using a POMDP formulation can significantly improve both safety and efficiency simultaneously. In particular, POMDP planning that includes state uncertainty is evaluated against both overconfident and conservative MDP models with only outcome uncertainty, showing that planning with state uncertainty has a benefit over both. The experimental method proposed here is also applicable to other cases in which internal-state-aware planning may improve performance.

The benefits of POMDP planning can only be realized with algorithms that can handle the continuous and irregular domains found in the real world. To that end, Chapter 4 proposes a pair of new algorithms for solving POMDPs with continuous state, action, and observation spaces. These algorithms are motivated by analysis and numerical experiments that show that leading online POMDP solvers cannot handle

continuous observation spaces. Previous solvers are shown to exhibit suboptimal behavior, explained by degenerate belief nodes with only a single state particle. One particular solver is proven to converge to the suboptimal QMDP approximation. The new algorithms, POMCPOW and PFT-DPW, handle this problem using progressive widening and weighted particle belief representations. Numerical experiments show that they are able to solve problems where previous methods fail.

Chapter 5 contains a description of the final contribution, which is a software package, POMDPs.jl, that aims to make the advances made in this thesis easier for others to build upon. The package uses the features of the Julia programming language to provide a convenient and flexible interface for expressing POMDPs. It also contains many tool for running simulations and has a suite of state-of-the art solvers and tools for writing new ones.

Much of the work in this thesis has been previously published in various venues [98, 99, 100, 32], but each section of the thesis contains important new additions.

Chapter 2

Trusted and Optimized UAV Collision Avoidance

This thesis's first investigation into safety and efficiency is in the context of unmanned aerial vehicles (UAVs). Collision avoidance is an important challenge that must be overcome to make UAV operation routine and safe. In this chapter, the suitability of the collision avoidance system for easy certification is the primary safety focus. The challenge is to create a system that is both easy to certify and has minimal impact on efficient normal operations.

2.1 Collision Avoidance for UAVs

As unmanned aerial vehicles (UAVs) move toward full autonomy, it is vital that they be capable of effectively responding to anomalous events, such as the intrusion of another aircraft into the vehicle's flight path. Minimizing collision risk for aircraft in general, and UAVs in particular, is challenging for a number of reasons. First, avoiding collision requires planning in a way that accounts for the large degree of uncertainty in the future paths of the aircraft. Second, the planning process must balance the competing goals of ensuring safety and avoiding disruption of normal operations. Many approaches have been proposed to address these challenges [57, 75, 5, 41, 42, 72, 102, 53, 54, 48, 7, 44, 80].

At present, there are two fundamentally different approaches to designing a conflict resolution system. The first approach focuses on inspiring confidence and trust in the system by making it as simple as possible for government regulators and vehicle operators to understand. This is accomplished by constructing the algorithm using hand-specified rules. Several algorithms that fit this paradigm have been proposed, and research toward formally verifying their safety-critical properties is underway [5, 41, 42, 72]. In this chapter, such algorithms are referred to as *trusted resolution logic* (TRL). Many TRL systems have a number of parameters that determine how conservatively the system behaves. These parameters may be tuned to meet specific performance goals, but their exact effects on performance can typically only be characterized empirically.

The second approach focuses on optimizing performance. This entails the offline or online computation of a “best” response action. Dynamic programming is widely used for this task [102, 53, 54, 48, 7, 44]. Conflict resolution systems designed using this approach will be referred to as *directly optimized* systems. Unfortunately, even if a conflict resolution system performs well in simulation, government regulators and vehicle operators will often (and sometimes rightly) be wary of trusting its safety due to perceived complexity and unpredictability. Even in the best case, such a system would require expensive and time-consuming development of tools for validation as was the case for the recently developed replacement for the traffic alert and collision avoidance system (TCAS) [44].

This chapter proposes two conflict resolution strategies that combine the strengths of trusted resolution logic and direct optimization approaches. In both strategies, dynamic programming is used to find near-optimal actions for each state. The difference lies in the set of actions available to the optimizer. In the *optimized TRL* approach, the actions are a set of certified TRL parameters. In the *trusted direct optimization* approach, the actions are a state-dependent set of immediate response actions that are deemed safe by certified TRL. Both strategies may achieve better performance than the base TRL without becoming significantly more difficult to certify.

The new approaches are tested in a scenario containing a UAV equipped with a perfect (noiseless) sensor to detect the state of an intruder and a simple reference

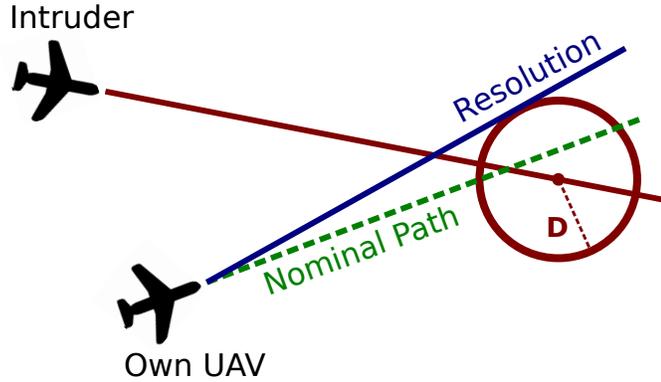


Figure 2.1: Trusted resolution logic. The near mid air collision exclusion zone (circle with radius D) moves with the intruder through time, but is shown here only at the time of closest approach. The TRL (Algorithm 1) finds a straight trajectory close to the nominal path that avoids this zone.

TRL to resolve conflicts. This TRL, illustrated in Figure 2.1, determines a path that does not pass within a specified separation distance, D , of the intruder given that the intruder maintains its current heading (except in cases where no such path exists or when the TRL’s heading resolution is too coarse to find such a path). This reference TRL is described in more detail in Section 2.2.4.

To account for uncertainty in the intruder’s flight path, if D takes a constant value, say \bar{D} , the value must be increased to ensure safety, but this may cause unnecessary departures from normal operation. To overcome this limitation and ensure safety without being too conservative, implementations of the two strategies mentioned above are applied. The optimized TRL approach computes a policy, π_D , that specifies a time varying separation distance, D_t , for each encounter state, while the direct optimization and trusted direct optimization approaches compute policies, π_ϕ and $\pi_{T\phi}$, that specify a bank angle, ϕ_t for each encounter state. For this reason, symbols referring to the optimized TRL approach generally have subscript D , while those referring to direct optimization and trusted direct optimization have a subscript ϕ .

The problem of dynamically selecting actions D_t or ϕ_t can be formulated as a Markov decision process (MDP). Online solution of π_D , π_ϕ , or $\pi_{T\phi}$ using an algorithm such as Monte Carlo Tree Search (MCTS) [23] would be conceptually straightforward,

but would require significant computing power onboard the vehicle, and it would be difficult and time-consuming to rigorously certify the implementation of MCTS due to its reliance on pseudo-random number generation. A key contribution of this chapter is to devise an offline approach to compute the policies. Offline optimization is difficult because of the size of the state space of the MDP, which is the Cartesian product of the continuous state spaces of the UAV and the intruder. This challenge is overcome with a value function approximation scheme that uses grid-based features that exploit the structure of the state space. This value function is optimized using simulation-based approximate value iteration.

2.2 MDP Models

The problem of avoiding an intruder using one of the strategies above is formulated as an MDP, referred to as an *encounter MDP*. An encounter involves two aerial vehicles flying in proximity. The first is the vehicle for which the resolution controller is being designed, which will be referred to as the “own UAV” (or simply “UAV”). The second is the intruder vehicle, which may be manned or unmanned and will be referred to as the “intruder”. Specifically, the encounter MDP is defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, R)$, which consists of

- The state space, \mathcal{S} : The state of the encounter, s , consists of (1) the state of the own UAV, $s^{(o)}$, (2) the state of the intruder, $s^{(i)}$, and (3) a boolean variable dev . The variable dev is set to true if the UAV has deviated from its nominal course. Collectively, the state is given by the triple

$$s = (s^{(o)}, s^{(i)}, \text{dev}). \quad (2.1)$$

The state components $s^{(o)}$ and $s^{(i)}$ are specified in Section 2.2.2. To model termination, the state space \mathcal{S} includes a “dummy” termination state, denoted with s_{term} .

- The action space, \mathcal{A} : The action space differs for the two approaches as described in detail in Section 2.2.5. a will be used to refer to all types of action

depending on the context.

- The state transition probability density function, $T : S \times A \times S \rightarrow \mathbb{R}$: The value $T(s, a, s')$ is the probability density of transitioning to state s' given that action a is taken at state s . This function is implicitly defined by a generative model that consists of a state transition function $F(\cdot)$ (described in Section 2.2.3) and a stochastic process W (described in Section 2.2.2).
- The reward, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$: The reward function, defined in Section 2.2.6, rewards reaching a goal and penalizes near mid air collisions (NMACs), deviation from the nominal path, and time outside a goal region.

2.2.1 Model Assumptions

Two important simplifying assumptions were made for this initial study. First, the UAV and the intruder move only in the horizontal plane and at constant speed. Modeling horizontal maneuvers is necessary because UAVs will likely have to employ horizontal maneuvers in place of or in addition to the vertical maneuvers that current collision avoidance systems for manned aircraft such as TCAS rely on. This is due to both climb performance limitations and regulatory constraints such as the 400ft ceiling for small (<55lb) commercial UAVs in Federal Aviation Administration rules [95]. Constraining altitude and speed simplifies exposition and reduces the size of the state space, keeping the computational burden modest compared with a formulation that includes variable altitude and speed. Extensions to higher fidelity models (e.g., [56]) are possible and left for future research. A higher fidelity model would present a computational challenge, but perhaps not an insurmountable one. For example, the TRL could be extended to handle variable speed and altitude, but the policy could be optimized only on the most important dimensions of the model (e.g., the horizontal plane). See [55] for a similar successful example.

Second, the intruder dynamics are independent of the UAV's state; in other words, the intruder does not react to the flight path of the UAV. The cooperative setting where both the UAV and the intruder are equipped with a collision avoidance system (CAS) [80, 105, 53] is left for future research.

2.2.2 Vehicle States and Dynamics

This paper uses a very simple discrete-time model of an encounter between two aerial vehicles with time steps of duration Δt . Throughout this section, the superscripts (o) and (i) refer to the own UAV and intruder quantities, respectively, while the superscript (\cdot) may be replaced by either (o) or (i) . Both the UAV's and the intruder's states consist of the horizontal position (x, y) and heading ψ , that is

$$s^{(o)} = (x^{(o)}, y^{(o)}, \psi^{(o)}), \quad s^{(i)} = (x^{(i)}, y^{(i)}, \psi^{(i)}). \quad (2.2)$$

The UAV and intruder also have similar dynamics. Both aircraft fly forward in the horizontal plane at constant speeds denoted by $v^{(\cdot)}$. They may turn at rates $\dot{\psi}^{(\cdot)}$ that remain constant over the simulation step. The following equations define the vehicle dynamics:

$$\begin{aligned} x_{t+1}^{(\cdot)} &= \begin{cases} x_t^{(\cdot)} + v^{(\cdot)} \cos(\psi_t^{(\cdot)}) \Delta t & \text{if } \dot{\psi}^{(\cdot)} = 0 \\ x_t^{(\cdot)} + v^{(\cdot)} \frac{\sin(\psi_t^{(\cdot)} + w_t \Delta t) - \sin(\psi_t^{(\cdot)})}{\dot{\psi}^{(\cdot)}} & \text{otherwise} \end{cases} \\ y_{t+1}^{(\cdot)} &= \begin{cases} y_t^{(\cdot)} + v^{(\cdot)} \sin(\psi_t^{(\cdot)}) \Delta t & \text{if } \dot{\psi}^{(\cdot)} = 0 \\ y_t^{(\cdot)} - v^{(\cdot)} \frac{\cos(\psi_t^{(\cdot)} + \dot{\psi}^{(\cdot)} \Delta t) - \cos(\psi_t^{(\cdot)})}{\dot{\psi}^{(\cdot)}} & \text{otherwise} \end{cases} \\ \psi_{t+1}^{(\cdot)} &= \psi_t^{(\cdot)} + \dot{\psi}^{(\cdot)} \Delta t. \end{aligned}$$

The intruder makes small random turns with

$$\dot{\psi}^{(i)} = w_t, \quad (2.3)$$

where w_t is a stochastic disturbance. We let W denote the stochastic process $\{w_t : t \in \mathbb{N}\}$. The random variables w_t in W are assumed independent and identically normally distributed with zero mean and a specified standard deviation, $\sigma_{\dot{\psi}}$. Subsequently, the intruder dynamics will be collectively referred to as $f^{(i)}$ so that

$$s_{t+1}^{(i)} = f^{(i)}(s_t^{(i)}, w_t). \quad (2.4)$$

The dynamics of the UAV are simplified conventional fixed wing aircraft dynamics with a single input, namely the roll angle $\phi^{(o)}$. We assume that the roll dynamics are fast compared to the other system dynamics, so that the roll angle $\phi^{(o)}$ may be directly and instantaneously commanded by the control system. This assumption avoids the inclusion of roll dynamics, which would increase the size of the state space.

$$\dot{\psi}_t^{(o)} = \frac{g \tan \phi_t^{(o)}}{v^{(o)}}, \quad (2.5)$$

where g is acceleration due to gravity. The performance of the UAV is limited by a maximum bank angle

$$|\phi_t^{(o)}| \leq \phi_{\max}. \quad (2.6)$$

The UAV dynamics will be collectively referred to as $f^{(o)}$, and

$$s_{t+1}^{(o)} = f^{(o)} \left(s_t^{(o)}, \phi_t^{(o)} \right). \quad (2.7)$$

2.2.3 Transition Function

It will often be convenient to refer to all of the dynamics in the state transition with a single transition function, which will be defined below after defining additional behavior regarding goals and near mid-air collisions.

The goal region that the UAV is trying to reach is denoted by S_{goal} . This is the set of all states in S for which $\left\| (x^{(o)}, y^{(o)}) - (x_{\text{goal}}^{(o)}, y_{\text{goal}}^{(o)}) \right\| \leq D_{\text{goal}}$, where $D_{\text{goal}} > 0$ is a specified goal region radius, and $(x_{\text{goal}}^{(o)}, y_{\text{goal}}^{(o)})$ is the goal center location. A near mid air collision (NMAC) occurs at time t if the UAV and intruder are within a minimum separation distance, $D_{\text{NMAC}} > 0$, that is if $\left\| (x_t^{(o)}, y_t^{(o)}) - (x_t^{(i)}, y_t^{(i)}) \right\| \leq D_{\text{NMAC}}$. If the UAV reaches the goal region at some time t , i.e., $s_t \in S_{\text{goal}}$, or if an NMAC occurs, the overall encounter state s transitions to the terminal state s_{term} and remains there. If the UAV performs a turn, `dev` is set to true because the vehicle has now deviated from the nominal straight path to the goal.

Let the state transition function, defined by (2.4), (2.7), and the special cases

above, be denoted concisely as F so that

$$s_{t+1} = F(s_t, \phi_t^{(o)}, w_t). \quad (2.8)$$

2.2.4 Reference Trusted Resolution Logic

The numerical tests use the simple reference TRL shown in Fig. 2.1. The TRL determines a heading angle, $\psi_{\text{resolution}}^{(o)}$, that is close to the heading to the goal, $\psi_{\text{goal}}^{(o)}$, and that will avoid future conflicts with the intruder given that the intruder maintains its current heading. This is shown in Figure 2.1.

Minimum Separation Distance

Given an initial state s , a candidate heading for the UAV, $\psi_{\text{cand}}^{(o)}$, and that both vehicles maintain their heading, the distance between the vehicles at the time of closest approach is a simple analytical function. Specifically, consider the distance $d(s, \psi_{\text{cand}}^{(o)}, \tau)$ between the vehicles τ time units in the future, i.e.,

$$d(s, \psi_{\text{cand}}^{(o)}, \tau) = \sqrt{\Delta x(\tau)^2 + \Delta y(\tau)^2}, \quad (2.9)$$

where

$$\begin{aligned} \Delta x(\tau) &= x^{(i)} - x^{(o)} + \tau v^{(i)} \cos(\psi^{(i)}) - \tau v^{(o)} \cos(\psi_{\text{cand}}^{(o)}), \\ \Delta y(\tau) &= y^{(i)} - y^{(o)} + \tau v^{(i)} \sin(\psi^{(i)}) - \tau v^{(o)} \sin(\psi_{\text{cand}}^{(o)}). \end{aligned}$$

The minimum distance between the two vehicles is analytically found by setting the time derivative of $d(s, \psi_{\text{cand}}^{(o)}, \tau)$ to zero. Specifically, the time at which the vehicles are closest is given by

$$\tau_{\min}(s, \psi_{\text{cand}}^{(o)}) = \max\left\{\frac{a+b}{c-2d}, 0\right\}, \quad (2.10)$$

where

$$\begin{aligned}
 a &:= -v^{(i)}x^{(i)} \cos(\psi^{(i)}) - v^{(i)}y^{(i)} \sin(\psi^{(i)}) \\
 b &:= v^{(o)}x^{(i)} \cos(\psi_{\text{cand}}^{(o)}) + v^{(o)}y^{(i)} \sin(\psi_{\text{cand}}^{(o)}) \\
 c &:= v^{(o)2} + v^{(i)2} \cos^2(\psi^{(i)}) + v^{(i)2} \sin^2(\psi^{(i)}) \\
 d &:= v^{(o)}v^{(i)}(\cos(\psi^{(i)}) \cos(\psi_{\text{cand}}^{(o)}) + \sin(\psi^{(i)}) \sin(\psi_{\text{cand}}^{(o)})).
 \end{aligned}$$

The minimum separation distance over all future time is then

$$d_{\min} \left(s, \psi_{\text{cand}}^{(o)} \right) := d \left(s, \psi_{\text{cand}}^{(o)}, \tau_{\min} \left(s, \psi_{\text{cand}}^{(o)} \right) \right). \quad (2.11)$$

It will also be occasionally convenient to use d_{\min} with only a state as an argument. In this case, $\psi_{\text{cand}}^{(o)}$ should be understood to be the UAV heading angle corresponding to that state, that is $d_{\min}(s) := d_{\min}(s, \psi^{(o)})$ where $\psi^{(o)}$ is the UAV heading in s .

Discrete Heading Optimization

The TRL begins with a discrete set of potential heading values (each denoted $\psi_{\text{cand}}^{(o)}$) for the UAV. It then determines, for a desired separation distance D , which of those will not result in a collision given that the UAV and intruder maintain their headings. Finally, it selects the value from that set which is closest to $\psi_{\text{goal}}^{(o)}$. The TRL is outlined in Algorithm 1.

2.2.5 Action Spaces and Control Systems

The numerical experiments consider policies generated by solving three different MDPs, each with a different action space. For the sake of conciseness, a will be used to represent any type of action depending on the context.

Direct Optimization

In the direct optimization approach, the actions are simply the bank angles less than ϕ_{\max} , that is $\mathcal{A}_{\phi} = \{\phi \in \mathbb{R} : |\phi| \leq \phi_{\max}\}$.

Algorithm 1 Trusted Resolution Logic

Input: Encounter state s , desired separation distance D
Output: Resolution heading angle $\psi_{\text{resolution}}^{(o)}$
function TRL(s, D)

 $\Psi \leftarrow \{\psi^{(o)} + n\pi/N : n \in \{-N, \dots, N\}\}$ ▷ range of values for heading
 $D^* = \max_{\psi_{\text{cand}}^{(o)} \in \Psi} d_{\min}(s, \psi_{\text{cand}}^{(o)})$
if $D^* < D$ ▷ conflict inescapable
 $\Psi \leftarrow \{\psi_{\text{cand}}^{(o)} \in \Psi : d_{\min}(s, \psi_{\text{cand}}^{(o)}) = D^*\}$
return $\underset{\psi_{\text{cand}}^{(o)} \in \Psi}{\text{argmin}} |\psi_{\text{cand}}^{(o)} - \psi_{\text{goal}}^{(o)}|$
else
 $\Psi \leftarrow \{\psi_{\text{cand}}^{(o)} \in \Psi : d_{\min}(s, \psi_{\text{cand}}^{(o)}) \geq D\}$
return $\underset{\psi_{\text{cand}}^{(o)} \in \Psi}{\text{argmin}} |\psi_{\text{cand}}^{(o)} - \psi_{\text{goal}}^{(o)}|$

Trusted Direct Optimization

In the trusted direct optimization approach, the set of actions is state dependent. Specifically, it is the set of bank angles that the TRL deems safe, that is $\mathcal{A}_{T\phi}(s) = \{\phi \in \mathcal{A}_\phi : d_{\min}(F(s, \phi, 0)) > D_{\text{NMAC}}\}$.

Optimized TRL

In the optimized TRL approach, the actions are the possible values for the separation distance, that is $\mathcal{A}_D = D \in \mathbb{R}_{\geq 0}$ used in the TRL (see Figure 2.1). A simple low level control system converts the desired heading from the TRL (denoted $\psi_{\text{resolution}}^{(o)}$) and into a bank angle for the vehicle. This is denoted with

$$\phi_t^{(o)} = c\left(s_t^{(o)}, \psi_{\text{resolution}}^{(o)}\right), \quad (2.12)$$

where $c(\cdot)$ represents the low level controller. When the action is a separation distance, the transition function, F , should be understood to contain the TRL and low level

control system, that is

$$F(s_t, D_t, w_t) := F(s_t, c(s_t^{(o)}, \text{TRL}(s_t, D_t)), w_t). \quad (2.13)$$

2.2.6 Reward

This section considers optimization of two competing metrics. The first goal is to minimize the risk of a NMAC. This aspect of performance is quantified with the fraction of NMACs prevented. The second goal is to minimize the probability of deviation from the nominal path. This metric was chosen (as opposed to a metric that penalizes the size of the deviation) because, in a commercial setting, any deviation from the normal operating plan might have a large cost in the form of disrupting schedules, preventing a mission from being completed, or requiring manual human monitoring. The MDP reward function is designed to encourage the policy towards a solution that performs well with respect to these goals.

Specifically, the utility associated with an encounter is the sum of the stage-wise rewards throughout the entire encounter

$$\sum_{t=0}^{\infty} \mathcal{R}(s_t, a_t). \quad (2.14)$$

In order to meet both goals, the stage-wise reward is

$$\begin{aligned} \mathcal{R}(s_t, a_t) := & -c_{\text{step}} + r_{\text{goal}} \times \text{in_goal}(s_t^{(o)}) \\ & - c_{\text{dev}} \times \text{causes_deviation}(s_t, a_t) \\ & - \lambda \times \text{is_NMAC}(s_t), \end{aligned} \quad (2.15)$$

for positive constants c_{step} , r_{goal} , c_{dev} , and λ . The first term is a small constant cost accumulated in each step to push the policy to quickly reach the goal. The function `in_goal` indicates that the UAV is within the goal region, so the second term is a reward for reaching the goal. The third term is a penalty for deviating from the nominal path. The function `causes_deviation` returns 1 if the action will cause

a deviation from the nominal course and 0 otherwise. It will only return 1 if the vehicle has not previously deviated and dev is false, so the penalty may only occur once during an encounter. This behavior makes the inclusion of dev in the state necessary. Constants c_{step} , r_{goal} , and c_{dev} represent relative weightings for the terms that incentivize a policy that reaches the goal quickly and minimizes the probability of deviation. Example values for these constants are given in Section 2.4. The fourth term is the cost for a collision. The weight λ balances the two performance goals. We heuristically expect there to be a value of λ for which the solution to the MDP meets the desired risk ratio if it is attainable. Bisection, or even a simple sweep of values can be used to find a suitable value, and this method has been used previously to analyze the performance of aircraft collision avoidance systems [7, 53].

2.2.7 Problem statement

The problem we consider is to find a feedback control policy $\pi^* : S \rightarrow A$, mapping an encounter state s_t into an action a_t , that maximizes the expected reward (2.14) subject to the system dynamics (2.8):

$$\begin{aligned} \underset{\pi}{\text{maximize}} \quad & E \left[\sum_{t=0}^{\infty} \mathcal{R}(s_t, \pi(s_t)) \right] \\ \text{subject to} \quad & s_{t+1} = F(s_t, \pi(s_t), w_t), \end{aligned} \tag{2.16}$$

for all initial states $s_0 \in S$. To make the solution of problem (2.16) practical, we present an approximate dynamic programming approach that yields a suboptimal policy, $\tilde{\pi}$, in the next section.

2.3 Solution Approach

Since the problem (2.16) has continuous state and action spaces and complex dynamics, it is difficult to solve. In order to make it more tractable, two approximations are used.

First, only a small number of discrete actions, which will be referred to as $\tilde{\mathcal{A}}$,

are considered. Specifically, for the optimized TRL, direct optimization, and trusted direct optimization approaches, the following action spaces are used:

$$\tilde{\mathcal{A}}_D = \{D_{\text{NMAC}}, 1.5D_{\text{NMAC}}, 2D_{\text{NMAC}}, 3D_{\text{NMAC}}, 4D_{\text{NMAC}}\} \quad (2.17)$$

$$\tilde{\mathcal{A}}_\phi = \left\{ -\phi_{\text{max}}, -\frac{\phi_{\text{max}}}{2}, 0, \frac{\phi_{\text{max}}}{2}, \phi_{\text{max}} \right\} \quad (2.18)$$

$$\tilde{\mathcal{A}}_{T\phi}(s) = \left\{ \phi \in \tilde{\mathcal{A}}_\phi : d_{\min}(F(s, \phi, 0)) > D_{\text{NMAC}} \right\}. \quad (2.19)$$

The second approximation is the approximate value iteration algorithm [10] used to optimize the policy. The value function, V , represents the expected value of the future reward given that the encounter is in state s and an optimal policy will be executed in the future. We approximate V with a linear architecture of the form

$$\tilde{V}(s) = \beta(s)^\top \theta, \quad (2.20)$$

where the feature function β returns a vector of N_β feature values, and $\theta \in \mathbb{R}^{N_\beta}$ is a vector of weights [10]. At each step of value iteration, the weight vector θ is fitted to the results of a large number of single-step simulations by solving a linear least-squares problem. After value iteration has converged, \tilde{V} is used to estimate a *post-decision state* value function, \tilde{V}_q , which is also approximated using a linear combination of features. The policy is extracted online in real time by selecting the action that results in the post-decision state that has the highest value according to \tilde{V}_q . The choice of working with post-decision states will be discussed in Section 2.3.2.

2.3.1 Approximate Value Iteration

The bulk of the computation is carried out offline before vehicle deployment using simulation. Specifically, the first step is to estimate the optimal value function for problem (2.16) using value iteration [10]. On a continuous state space such as the encounter state space used in this work, the Bellman operator used in value iteration cannot be applied for each of the uncountably infinite number of states, so an approximation must be used.

In this paper we adopt projected value iteration [10]. In this algorithm, each successive value function approximation, \tilde{V}_k , is the result of the Bellman operation projected onto a linear subspace with respect to the Euclidean norm, that is

$$\tilde{V}_{k+1}(s) = \Pi \mathcal{B}[\tilde{V}_k](s), \quad (2.21)$$

where \mathcal{B} is the Bellman operator, and Π is a projection onto the linear subspace Φ spanned by the N_β basis functions (see [10] for a detailed discussion of this approach).

Two important factors that make projected value iteration an attractive option compared to other approaches, in particular deep reinforcement learning [70], are the following: First, since it is relatively easy to choose a feature set that yields acceptable performance by hand (see Section 2.3.4), more flexible function approximation approaches that also learn features, such as deep neural networks [39], are not necessary. Second, the least squares optimization problem solved at each iteration of projected value iteration (2.23) requires much less computation than the gradient descent required to train a neural network.

Monte Carlo simulations are used to perform the approximate value iteration (2.21). Specifically, for each iteration, N_{state} states are uniformly randomly selected. If the states lie within the grids used in the feature function (see Section 2.3.4) the sample is “snapped” to the nearest grid point to prevent approximation errors due to the Gibbs phenomenon [33]. At each sampled state $s^{[n]}$, $n = 1, \dots, N_{\text{state}}$, the stage-wise reward and the expected value of the value function are evaluated for each action a in $\tilde{\mathcal{A}}$. The expectation embedded in the Bellman operator is approximated using N_{EV} *single-step* intruder simulations, each with a randomly generated noise value, w_m , $m = 1, \dots, N_{\text{EV}}$. However, since the own UAV dynamics are deterministic, only one own UAV simulation is needed. The maximum over $\tilde{\mathcal{A}}$ is selected and stored as the n th entry of a vector v_{k+1} :

$$v_{k+1}[n] := \max_{a \in \tilde{\mathcal{A}}} \left\{ \mathcal{R}(s^{[n]}, a) + \frac{1}{N_{\text{EV}}} \sum_{m=1}^{N_{\text{EV}}} \beta (F(s^{[n]}, a, w_m))^\top \theta_k \right\}, \quad (2.22)$$

for $n = 1, \dots, N_{\text{state}}$. Here v_{k+1} provides an approximation to the (unprojected) value

function. To project v_{k+1} onto Φ , we compute the weight vector θ_{k+1} by solving the least-squares optimization problem

$$\theta_{k+1} = \underset{\theta \in \mathbb{R}^{N_\beta}}{\operatorname{argmin}} \sum_{n=1}^{N_{\text{state}}} \left(\beta (s^{[n]})^\top \theta - v_{k+1}[n] \right)^2. \quad (2.23)$$

Iteration is terminated after a fixed number of steps, N_{VI} , and the resulting weight vector, denoted θ , is stored for the next processing step (Section 2.3.2).

2.3.2 Post Decision Value Function Extraction

For reasons discussed in Section 2.3.3, the second step is to approximate a value function defined over post-decision states [10]. A post-decision state, denoted by q , is defined as a state in S consisting of the own UAV state and dev *at one time step into the future* and the intruder state *at the current time*, that is

$$q_t = \left(s_{t+1}^{(o)}, s_t^{(i)}, \operatorname{dev}_{t+1} \right). \quad (2.24)$$

Correspondingly, let $g : S \times A \rightarrow S$ be the function that maps the current state and action to the post-decision state. In other words, function $g(s_t, a_t)$ returns q_t consisting of

$$\begin{aligned} s_{t+1}^{(o)} &= f^{(o)} \left(s_t^{(o)}, a_t \right) \\ s_t^{(i)} &= s_t^{(i)} \\ \operatorname{dev}_{t+1} &= \max\{\operatorname{dev}, \operatorname{causes_deviation}(s_t, a_t)\}. \end{aligned} \quad (2.25)$$

The approximate value function over post-decision states, \tilde{V}_q , is computed as follows. Let $h : S \times \mathbb{R} \rightarrow S$ be a function that returns the next encounter state given the post decision state and intruder heading noise value, that is $h(q_t, w_t)$ returns s_{t+1}

consisting of

$$s_{t+1}^{(o)} = s_{t+1}^{(o)} \quad (2.26)$$

$$s_{t+1}^{(i)} = f^{(i)}(s_t^{(i)}, w_t) \quad (2.27)$$

$$\text{dev}_{t+1} = \text{dev}_{t+1}, \quad (2.28)$$

where $(s_{t+1}^{(o)}, s_t^{(i)}, \text{dev}_{t+1})$ are the members of q_t .

The value function \tilde{V}_q can then be expressed in terms of \tilde{V} as

$$\tilde{V}_q(q) = E_{w_t} \left[\tilde{V}(h(q, w_t)) \right], \quad (2.29)$$

where w_t denotes, as usual, a random variable with Gaussian normal density. Equation (2.29) implies that \tilde{V}_q can also be approximated according to a linear architecture

$$\tilde{V}_q(q) = \beta(q)^\top \theta^q, \quad (2.30)$$

where $\beta(q)$ is the feature vector for post-decision states $q \in S$, and $\theta^q \in \mathbb{R}^{N_\beta}$ is the corresponding weight vector.

Specifically, N_q post decision states are randomly selected using the same method as described in Section 2.3.1 and are denoted as $q^{[n]}$, $n = 1, \dots, N_q$. For each sampled state $q^{[n]}$, the expectation in (2.29) is approximated using N_{EV} single-step simulations. The results are used to solve a least squares optimization problem

$$\theta^q = \underset{\theta \in \mathbb{R}^{N_\beta}}{\text{argmin}} \sum_{n=1}^{N_q} \left(\beta(q^{[n]})^\top \theta - v^q[n] \right)^2, \quad (2.31)$$

where

$$v^q[n] := \frac{1}{N_{\text{EV}}} \sum_{m=1}^{N_{\text{EV}}} \beta(h(q^{[n]}, w_m))^\top \theta, \quad (2.32)$$

where w_m , $m = 1, \dots, N_{\text{EV}}$, is a noise value sampled from the distribution of a random variable in W .

2.3.3 Online Policy Evaluation

The first two steps (explained, respectively, in Sections 2.3.1 and 2.3.2) are performed offline. The last step, namely policy evaluation, is performed online. Specifically a suboptimal control at any state s is computed as

$$\tilde{\pi}(s) = \operatorname{argmax}_{a \in \tilde{A}} \tilde{V}_q(g(s, a)). \quad (2.33)$$

Since g (defined in (2.25)) is a *deterministic* function of a state-action pair, this calculation does not contain any computationally costly or difficult-to-certify operations such as estimating an expectation.

Two comments regarding the motivation for using the post-decision value function are in order. First, if the value functions could be exactly calculated, the post-decision state approach would be equivalent to the more common approach wherein a control is computed by solving

$$\pi(s) = \operatorname{argmax}_{a \in \tilde{A}} Q(s, a), \quad (2.34)$$

In fact, one can readily show $Q(s, a) = V_q(g(s, a))$. However, when approximations are used, post-decision states provide a more robust (i.e., less susceptible to approximation error) way of deriving control actions [10]. One reason for this is that the cost function is approximated in the space of post-decision states, rather than in the larger space of state-control pairs, and hence the post-decision method is less susceptible to complications with inadequate exploration [10].

Second, the post decision value function is not used for the value iteration portion of the offline solution as it would require full simulations of both the own UAV and the intruder dynamics and, therefore, would be more computationally demanding.

2.3.4 Selection of Features

The primary value function approximation features are the interpolation weights for points in a grid [24]. A grid-based approximation is potentially inefficient compared

to a small number of global features (e.g. heading, distance, and trigonometric functions of the variables), however, it is well known that the function approximation used in value iteration, must have suitable convergence properties [13] in addition to approximating the final value function. Indeed, experiments with a small number of global features were attempted, but they did not achieve convergence, and a grid approach was adopted instead. Since a grid defined over the entire six dimensional encounter state space with a reasonable resolution would require far too many points to be computationally feasible, the grid must be focused on important parts of the state space.

Our strategy is to separate features into two groups (along with a constant), specifically

$$\beta(s) = [\beta_{\text{intruder}}(s^{(o)} - s^{(i)}), \beta_{\text{goal}}(s^{(o)}), 1]. \quad (2.35)$$

The first group, β_{intruder} , captures the features corresponding to a near midair collision and is a function of only the position and orientation of the own vehicle relative to the intruder. The second group, β_{goal} , captures the value of being near the goal and is a function of only the own vehicle state.

Since the domain of β_{intruder} is only three dimensional and the domain of β_{goal} is only two dimensional, relatively fine interpolation grids can be used for value function approximation without requiring a prohibitively large number of features. The β_{intruder} feature group consists of a NMAC indicator function and interpolation weights for a grid (Figure 2.2) with nodes at regularly spaced points along the following three variables: (1) the distance between the UAV and intruder, (2) the bearing from the intruder to the UAV, and (3) the relative heading between the vehicles. The β_{goal} vector consists of a goal indicator function, the distance between the UAV and the goal, and interpolation weights for a grid (Figure 2.3) with nodes regularly spaced along the distance between the UAV and the goal and the absolute value of the bearing to the goal from the UAV. The total number of features is $N_\beta = 1813$.

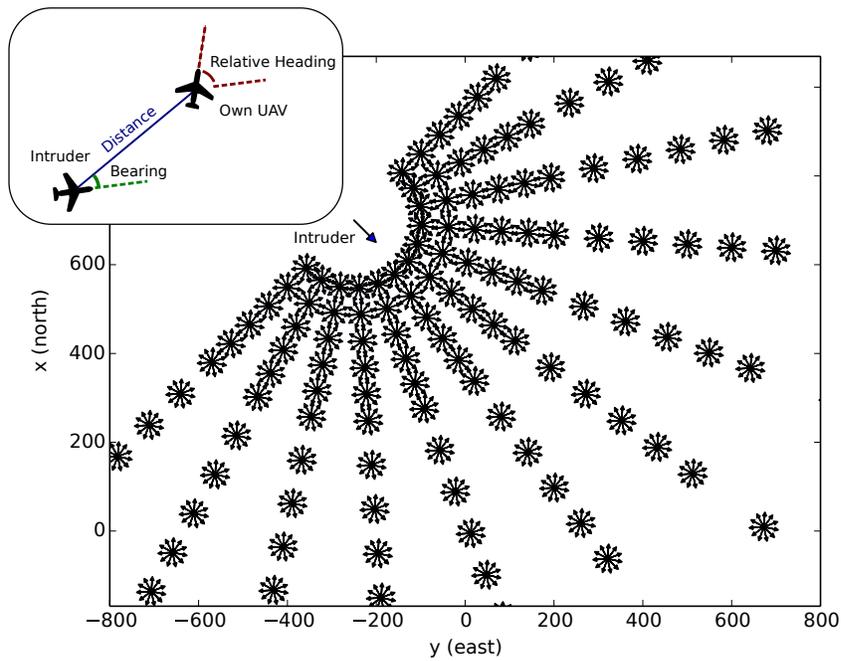


Figure 2.2: Intruder interpolation grid for β_{intruder} , visualized when the intruder is located at (700 m, -250 m) at heading 135. The top left inset shows the variables used. In the main plot, each of the small arrows represents a grid point. At each of the point locations, there are twelve small arrows radiating out. Each arrow represents a different UAV heading.

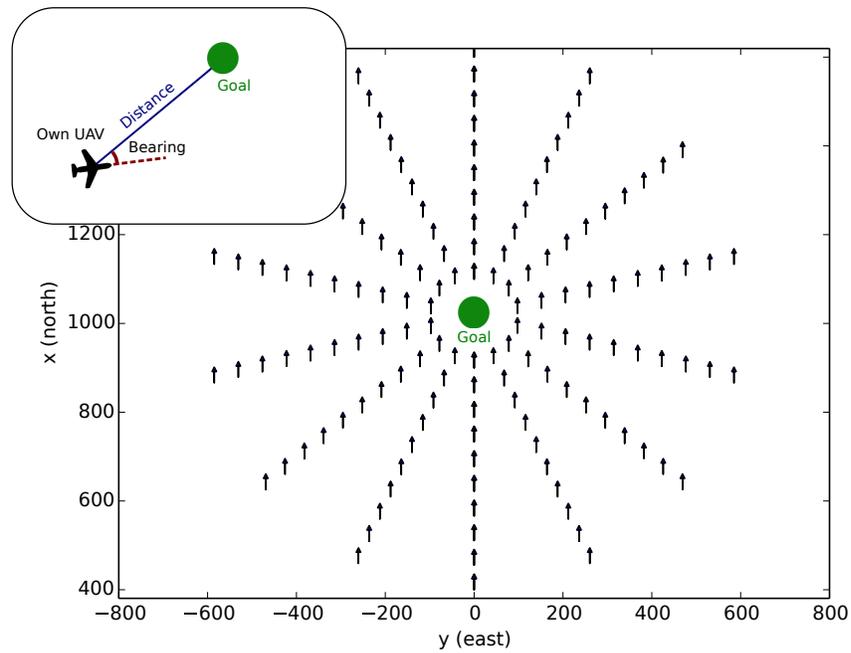


Figure 2.3: Goal interpolation grid for β_{goal} when the own UAV's heading is directly north. The grid takes advantage of symmetry in the bearing variable, so, for each arrow on the left side, there is an arrow on the right side that corresponds with the *same* point in the grid.

2.4 Results

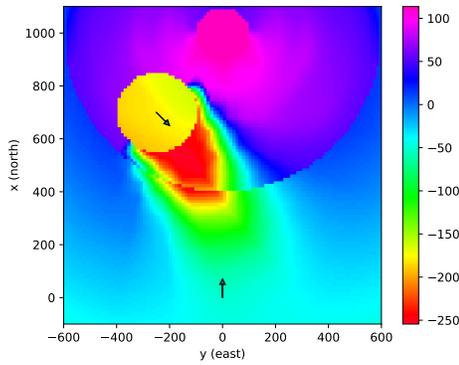
This section presents results from numerical experiments that illustrate the effectiveness of the new approach. The experiments are designed to compare the four approaches discussed in Section 2.1: the “static TRL” approach, “direct optimization”, and the new “optimized TRL” and “trusted direct optimization” approaches proposed in this chapter. The static TRL control law uses the TRL described in Algorithm 1 with a constant value for the separation distance (denoted with \bar{D}), while the other approaches use the approximate optimization procedure described in Section 2.3. The various parameters used in the numerical experiments are listed in Table 2.1.

2.4.1 Policies

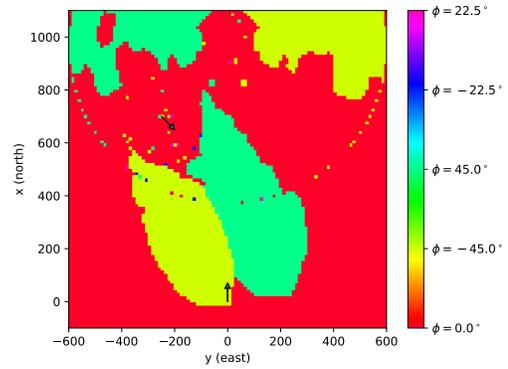
This section includes visualizations of two-dimensional “slices” of several UAV policies and their associated value functions. In each of the slices, the intruder is located at (700 m, -250 m) pointed at heading 135 as indicated by the arrow. The goal is at (1000 m, 0). Each pixel on this image represents the value function or policy evaluated with the UAV at that position pointed directly north. As expected, for all approaches, there is a low value region in front of and to the south of the intruder and a high value region near the goal.

Figures 2.4a and 2.4b show the value function and policy for the directly optimized approach, and Figs. 2.4c and 2.4d show the same for the trusted directly optimized approach. The policies for both approaches are similar, with regions in front of the intruder where sharp turns are commanded. The trusted direct optimization policy is slightly more conservative with larger turn regions because of the restricted action space.

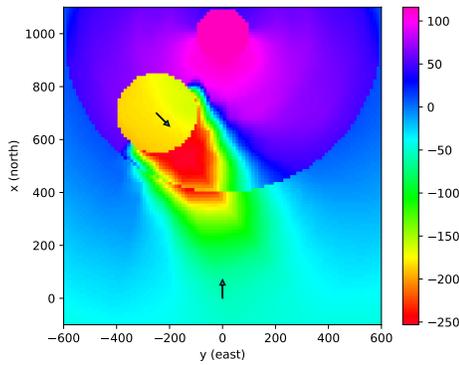
Figure 2.4f shows the value function and policy for the optimized TRL approach. When multiple actions result in the same post decision state value, the least conservative action is chosen, so the policy yields the lowest value of D (500ft \approx 152.4m) on most of the state space. Because the own UAV is pointed north, the policy is



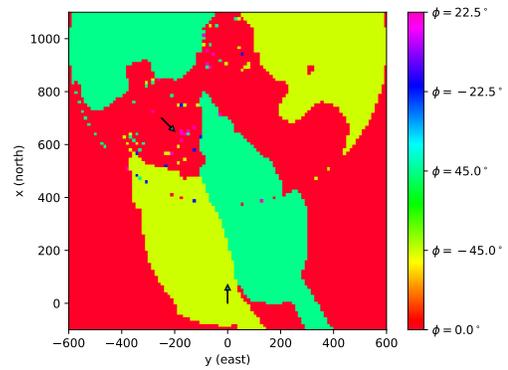
(a) Slice of the approximate optimal value function for direct optimization.



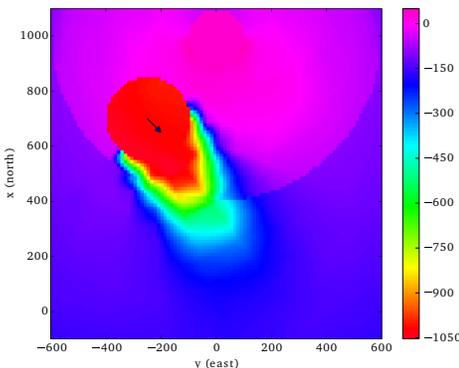
(b) Slice of the direct optimization policy.



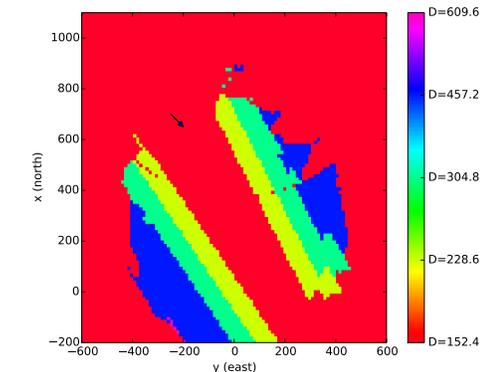
(c) Slice of the approximate optimal value function for trusted direct optimization.



(d) Slice of the trusted direct optimization policy.



(e) Slice of the approximate optimal value function for the optimized TRL.



(f) Slice of the optimized TRL policy.

Figure 2.4: UAV Collision avoidance policy visualizations

conservative in a region in front of and to the south of the intruder. The band corresponding to small D that stretches across the middle of the conservative region (from (700 m, -250 m) to (-100 m, 200 m)) is present because all values of D result in the same post decision value.

2.4.2 Numerical Performance Evaluation

The control policies are evaluated by executing them in a large number of complete (from $t = 0$ to the end state) encounter simulations. The same random numbers used to generate intruder noise were reused across all collision avoidance strategies to ensure fairness of comparisons. In each of the simulations, the own UAV starts pointed north at position (0, 0) in a north-east coordinate system with the goal at (1000m, 0).

The evaluation simulations use the same intruder random turn rate model with standard deviation σ_ψ that was used for value iteration. A robustness study using different models is not presented here, but previous research [55] suggests that this method will offer good performance when evaluated against both a range of noise parameters and structurally different models. The intruder initial position is randomly generated between 800 m and 1500 m from the center point of the encounter area at (500m, 500m) with an initial heading that is within 135° of the direction from the initial position to the center point.

The conservativeness of each control law is characterized by counting the number of deviations from the nominal path in 10,000 simulations with initial conditions shown in Fig. 2.5. Of these simulations, 1009 result in a NMAC if the UAV follows its nominal path, but for most a deviation would not be necessary to avoid the intruder.

The fraction of collisions avoided is estimated using a separate set of 10,000 simulations. Each of these simulations has an initial condition in the same region described above, but initial conditions and noise trajectories are chosen by filtering random trials so that each of the simulations *will result in a NMAC if the own UAV follows its nominal path*.

Figure 2.6 shows the Pareto optimal frontiers for the different control laws. Each

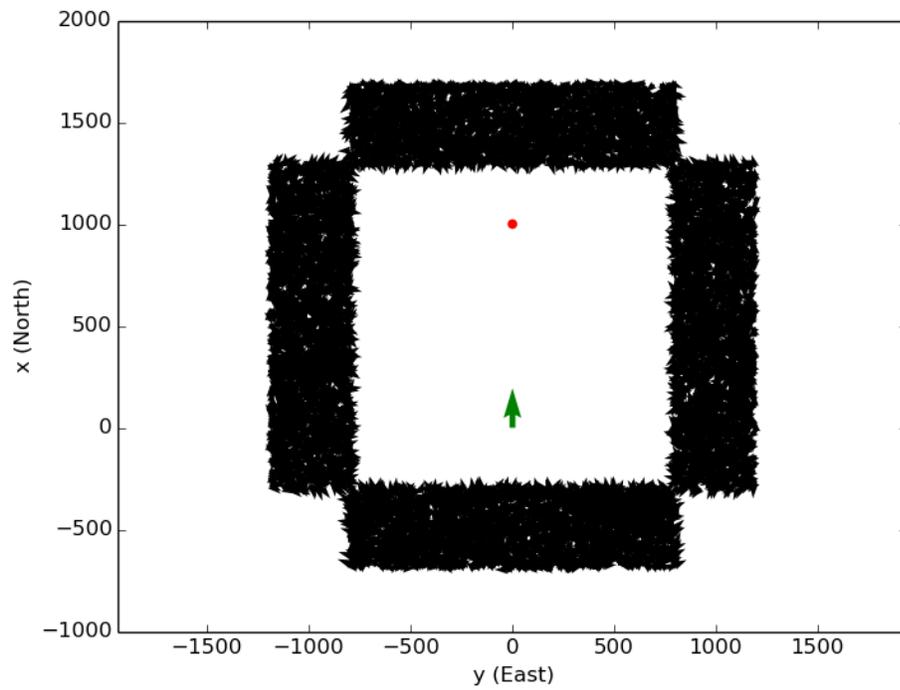


Figure 2.5: Intruder initial conditions for evaluation simulations. Each small black arrow is an intruder initial state. The large green arrow is the UAV's initial state. The red dot is the goal.

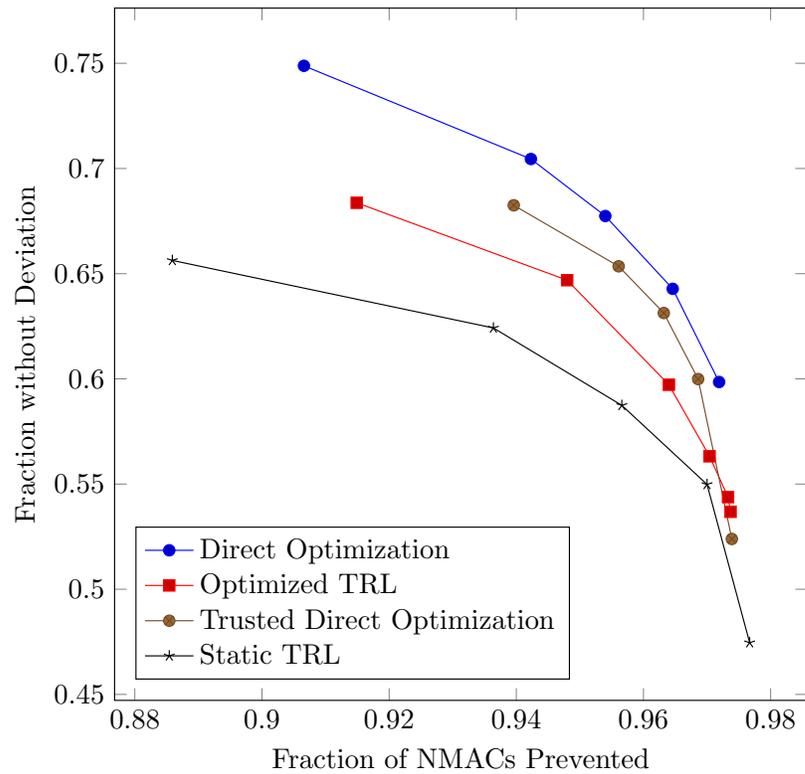


Figure 2.6: Policy performance comparison. The values of λ used to generate the datapoints are 100, 316, 1000, 3160, 10^4 , and 3.16×10^4 for the optimized TRL policy and 300, 500, 700, 1000, and 1500 for the directly optimized approach. The values of \bar{D} for the static TRL policy are 250 m, 300 m, 350 m, 400 m, and 500 m.

Table 2.1: Parameters for numerical experiments

Description	Symbol	Value
Own UAV speed	$v^{(o)}$	30 m/s
Maximum own UAV bank angle	ϕ_{\max}	45°
Intruder speed	$v^{(i)}$	60 m/s
Intruder turn rate standard deviation	$\sigma_{\dot{\psi}}$	10°/s
Near mid air collision radius	D_{NMAC}	500 ft
Step cost	c_{step}	1
Reward for reaching goal	r_{goal}	100
Cost for deviation	c_{dev}	100
Step simulations for expectation estimate	N_{EV}	20
Single step simulations per round of value iteration (optimized TRL)	N_{state}	10,000
Single step simulations per round of value iteration (directly optimized)	N_{state}	50,000
Number of value iteration rounds	N_{VI}	35
Single step simulations for post decision value func- tion extraction	N_q	50,000

curve is generated by using various values of λ in the reward function of the MDP, or by using various static values of \bar{D} in the static TRL case. It is clear that the optimization provides better performance than the static TRL. For example, if the desired fraction of NMACs prevented is 96%, interpolation between data points suggests that direct optimization will cause approximately 20% fewer deviations than the static TRL policy. This difference may be interpreted as the *price* if using trusted resolution logic rather than optimization.

Fortunately, both the optimized TRL and trusted direct optimization approaches offer ways to reduce this price. Neither should be much more difficult to certify than the TRL because they are both based on the TRL. In particular, neither will ever command an action that is deemed unsafe by the TRL. Thus, they provide the same level of trust as the TRL and thus reduce the price of trust. In this case, trusted direct optimization is able to nearly close the gap between static TRL and direct optimization, while optimized TRL reduces the gap by about half, so trusted direct optimization would be the preferred approach.

The software used to simulate these experiments was written in the Julia programming language, and is freely available at <https://github.com/zsunberg/UASEncounter>.

2.5 Discussion

This chapter has explored using trusted resolution logic alongside approximate optimization to produce a high performance collision avoidance system for UAV that is also relatively easy to certify. Due to flight performance and regulatory constraints, only horizontal maneuvers are considered when resolving potential collisions and the intruding aircraft is presumed to be unaware of the UAV. This problem is formulated as a Markov decision process. It is solved using an approximate value iteration approach. This general approach is widely used [52], though problems of this size are still challenging to handle. By choosing appropriate features and using a post-decision state, the method is made tractable for this problem.

Using simple trusted resolution logic that guarantees a separation distance D if the intruder maintains its current heading is an alternate approach. This approach has the advantage that it can be easily certified because, by construction, it will never command a heading that is considered dangerous. However, when this TRL is compared to an approximate dynamic programming solution to the MDP, it is too conservative, causing unnecessary disruption to normal flight. These tests demonstrate that there is a price that comes with certifiability and trustworthiness, and the performance gap between the approaches quantifies this price. Two alternative combinations of approximate dynamic programming and the TRL are considered as ways to reduce the price.

First, approximate dynamic programming is applied to dynamically choose D within a certified range based on the encounter state. Since every potential action is pre-certified, this approach retains the ease of certification of the TRL approach. However, since this new action space is relatively limited, it only reduces the price by about half.

The second approach is to filter the original action space of turning maneuvers

through the TRL so that the UAV may only take actions that will maintain D separation based on the current state. This approach should also be easy to certify, and it largely closes the performance gap, nearly eliminating the price of certifiability.

There are many potential extensions for this work. First, the function approximation for the value iteration could be improved. The feature-based approach used here worked well, but the selection of the features was mostly heuristically justified, and another more advanced and general approximation technique such as neural networks [39] may work better. Most importantly, the uncertainty model should be improved. The current model was not justified with data, and it is extremely unlikely that a Gaussian noise model reflects real intruder behavior. It is much more plausible that the intruder has an operational intention unknown to the UAV which should be modeled as a hidden state in the problem. It is not immediately clear how to solve the resulting POMDP offline, but a QMDP approach using the MDP solution found in this work may be adequate. This concept of a hidden internal state governing the behavior of other agents is investigated in the next chapter, but in a different context: self-driving cars.

Chapter 3

Planning with Internal States in Driving

The second investigation of this thesis is in the area of autonomous driving. Chapter 2 focuses on finding the best way to integrate certifiable safety constraints, but uses a simplistic stochastic model for the intruder the UAV interacts with. This chapter keeps the safety constraint architecture discovered in the last chapter as a constraint on the action space that guarantees there will be no crashes (Section 3.2.3), but the chapter focuses primarily on investigating a richer stochastic model. In particular, it demonstrates the advantage of using a model with other agents' internal states represented using state uncertainty in the problem formulation.

3.1 Human Interaction in Autonomous Driving

One challenge in introducing autonomous automobiles is ensuring that they interact safely with human drivers. In order to navigate complex driving scenarios, human drivers routinely predict what other drivers will do and make driving decisions based on these predictions. Autonomous vehicles typically take an overly conservative approach, which can result in physical danger, reduced efficiency, and an uncomfortable experience. In a 2015 study, autonomous vehicles drove over 1.2 million miles without

being legally responsible for any accidents. However, the autonomous vehicles actually had a higher accident rate than average for a conventional vehicle in the United States because of accidents for which they were not legally responsible [89]. This result suggests that there is significant room for improvement in autonomous-human vehicle interaction.

One approach to improve interaction would be to program ad-hoc logic for each situation into the vehicles. However, this approach is time-consuming and error prone, and edge cases that the programmers have not foreseen can present a safety risk. Furthermore, this approach limits the performance of the system to the capability of the human programmer. In contrast, artificial intelligence and machine learning techniques have the potential to provide a more robust approach to such decision-making tasks. This chapter explores MDP and POMDP techniques.

POMDPs are particularly well suited for modeling decisions for autonomous vehicles because they explicitly capture the limitations of the vehicle's sensors in measuring the relevant state variables [15, 84, 6]. Though sensors can accurately measure many of the relevant variables pertaining to the physical state of the vehicles, the internal state (e.g., intentions and aggressiveness) of other drivers and road users can only be indirectly inferred [84, 6, 61, 29]. The hypothesis explored in this chapter is that inferring and planning with an estimate of the internal states of the traffic participants will improve safety and efficiency.

Driving strategies derived from MDPs and POMDPs depend on several ingredients to be successful. First, an accurate stochastic model of the environment, including the behavior of other drivers, is necessary. Though this chapter uses a very simple model, there has been significant work on creating better models in recent years [36, 110, 82], and the POMDP planning methods used here can be adapted to use these new models.

Before investing the effort required to develop and test a POMDP-based decision making system for real autonomous vehicles, it is important to quantify the potential performance improvement. This chapter presents a method that involves comparing solutions obtained from several variations of Monte Carlo Tree Search [17]. For this research, these ideas are investigated in the context of making lane changes on a

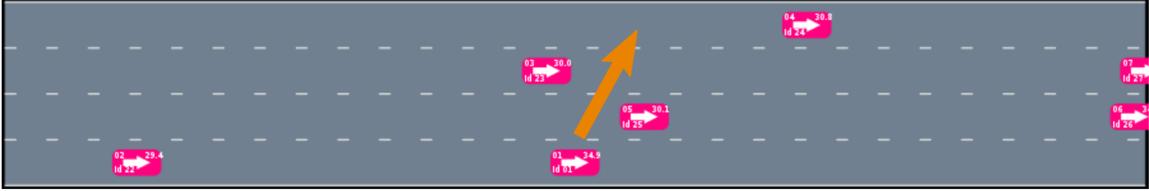


Figure 3.1: The example decision making scenario for this chapter. An autonomous vehicle (bottom center) must travel from the rightmost to leftmost lane within a limited distance.

freeway. This situation that has been anecdotally noted to be difficult [73], and a good candidate for improvement by reasoning about the actions of other drivers. For instance, a 2016 Google report noted that “merging into traffic during rush hour is an exercise in negotiation.” [26].

The experiments in this chapter are conducted as follows: First, as a baseline for evaluation, the performance gains that would result from perfect estimation of and planning with internal states are estimated. Next, simulations evaluate whether a particle filter can be used to estimate internal parameters using physical measurements. Finally, the performance gain from using QMDP and POMDP problem formulations to plan with these internal state estimates is assessed.

3.2 Freeway Driving POMDP

The focus of this chapter is on freeway driving. In the scenario studied here, a vehicle must navigate from the rightmost to the leftmost lane of a four lane freeway within a specified distance while maintaining safety and comfort (see Fig. 3.1).

Throughout this section, x denotes position in the *longitudinal* direction, that is, the direction that the cars move along the road in meters, and y denotes position in the *lateral* direction, that is, the lane the car occupies in lane units. The problem can be stated as a discrete-time POMDP defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \mathcal{O}, \mathcal{Z})$, which consists of

- The state space, \mathcal{S} : A system state,

$$s = (q_0, \{(q_i, \theta_i)\}_{i \in 1..N}) \in \mathcal{S},$$

consists of the physical state of the ego vehicle (q_0), and physical state and behavior model for each of the N other cars in the scene. The physical state,

$$q_i = (x_i, y_i, \dot{x}_i, \dot{y}_i),$$

consists of the car's longitudinal and lateral position and velocity. The internal state (behavior model parameters), θ_i , is drawn from a set of behaviors Θ .

- The action space, \mathcal{A} : An action, $u = (\ddot{x}_e, \ddot{y}_e) \in \mathcal{A}$, consists of the longitudinal acceleration and lateral velocity of the ego vehicle. The action space is discrete and pruned to prevent crashes (see Section 3.2.3).
- The state transition model, $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$: The value $\mathcal{T}(s, u, s')$ is the probability of transitioning to state s' given that action u is taken by the ego at state s . This function is implicitly defined by a generative model that consists of a state transition function, $F(\cdot)$, and a stochastic noise process (see Section 3.2.2).
- The reward model, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$: The reward function, defined in Section 3.2.4, rewards reaching the left lane within the distance limit and penalizes unsafe actions.
- The observation space, \mathcal{O} : An observation, $o \in \mathcal{O}$ consists of the physical states of all of the vehicles, that is $o = \{q_i\}_{i \in 1..N}$. No information about the internal state is directly included in the observation.
- The observation model, $Z : \mathcal{S} \times \mathcal{O} \rightarrow \mathbb{R}$: The value $Z(s', o)$ is the probability of receiving observation o when the system transitions to state s' . In these experiments, the physical state is assumed to be known exactly, though it is not difficult to relax this assumption.

The remainder of this section elaborates on this model.

3.2.1 Driver Modeling

The driver models for each car have two components: an acceleration model that governs the longitudinal motion and a lane change model that determines the lateral motion. In this chapter, the acceleration model is the Intelligent Driver Model (IDM) [106], and the lane change model is the “Minimizing Overall Braking Induced by Lane change” (MOBIL) model [50]. Both of these models have a small number of parameters that determine the behavior of the drivers. The distribution of these parameters in the population of vehicles will be denoted Θ .

IDM

The IDM Model was developed as a simple model for “microscopic” simulations of traffic flows and is able to reproduce some phenomena observed in real-world traffic flows [106]. It determines the longitudinal acceleration for a human-driven car, \ddot{x} , based on the desired distance gap to the preceding car, g , the absolute velocity, \dot{x} , and the velocity relative to the preceding car $\Delta\dot{x}$. The longitudinal acceleration is governed by the following equation:

$$\ddot{x}_{\text{IDM}} = a \left[1 - \left(\frac{\dot{x}}{\dot{x}_0} \right)^\delta - \left(\frac{g^*(\dot{x}, \Delta\dot{x})}{g} \right)^2 \right], \quad (3.1)$$

where g^* is the desired gap given by

$$g^*(\dot{x}, \Delta\dot{x}) = g_0 + T\dot{x} + \frac{\dot{x}\Delta\dot{x}}{2\sqrt{ab}}. \quad (3.2)$$

Brief descriptions and values for the parameters not defined here are provided later in Table 3.1.

A small amount of noise is also added to the acceleration

$$\ddot{x} = \ddot{x}_{\text{IDM}} + w, \quad (3.3)$$

where w is a random variable with a triangular distribution with support between $-a/2$ and $a/2$. In cases where the noise might cause a hard brake or lead to a state where a crash is unavoidable, the distribution is scaled appropriately.

MOBIL

The MOBIL model makes the decision to change lanes based on maximizing the acceleration for the vehicle and its neighbors. When considering a lane change, MOBIL first ensures that the safety criterion $\tilde{\ddot{x}}_n \geq -b_{\text{safe}}$, where $\tilde{\ddot{x}}_n$ will be the acceleration of the following car if the lane change is made and b_{safe} is the safe braking limit. It then makes the lane change if the following condition is met

$$\tilde{\ddot{x}}_c - \ddot{x}_c + p(\tilde{\ddot{x}}_n - \ddot{x}_n + \tilde{\ddot{x}}_o - \ddot{x}_o) > \Delta a_{\text{th}} \quad (3.4)$$

where the quantities with tildes are calculated assuming that a lane change is made, the quantities with subscript c are quantities for the car making the lane change decision, those with n are for the new follower, and those with o are for the old follower. The parameter $p \in [0, 1]$ is the politeness factor, which represents how much the driver values allowing other vehicles to increase their acceleration. The parameter Δa_{th} is the threshold acceleration increase to initiate a lane changing maneuver. Parameter values are listed in Table 3.1.

3.2.2 Physical Dynamics

The physical dynamics are simplified for the sake of computational efficiency. Time is divided into discrete steps of length Δt . The longitudinal dynamics assume constant acceleration, and the lateral dynamics assume constant velocity over a time step, that is

$$\begin{aligned} x' &= x + \dot{x}\Delta t + \frac{1}{2}\ddot{x}\Delta t^2 \\ \dot{x}' &= \dot{x} + \ddot{x}\Delta t \\ y' &= y + \dot{y}\Delta t. \end{aligned}$$

There is a physical limit to the braking acceleration, b_{\max} . Lateral velocity is allowed to change instantly because cars on a freeway can achieve the lateral velocity needed for a lane change in time much shorter than Δt by steering. If MOBIL determines that a lane change should be made, the lateral velocity, \dot{y} , is set to \dot{y}_{lc} . Lane changes are not allowed to reverse. Once a lane change has begun, \dot{y} remains constant until the lane change is completed (this is the reason that \dot{y} is part of the state). When a vehicle passes over the midpoint of a lane, lateral movement is immediately stopped so that lane changes always end at exactly the center of a lane.

Since MOBIL only considers cars in adjacent lanes, there must be a coordination mechanism so that two cars do not converge into the same lane simultaneously. In order to accomplish this, if two cars begin changing into the same lane simultaneously, and the front vehicle is within g^* of the rear vehicle, the rear vehicle's lane change is canceled.

In order to reduce the computational demands of decision-making, only 50 m of road in front of the ego and 50 m behind are modeled. Thus, a model for vehicle entry into this section is needed. If there are fewer than N_{\max} vehicles on the road, a new vehicle is generated. First, a behavior for the new vehicle is drawn from Θ , and the initial speed is set to $\dot{x}_0 + \sigma_{\text{vel}}w_0$, where \dot{x}_0 is the desired speed from the behavior model and w_0 is a zero-mean, unit-variance, normally distributed random variable that is independent for each car. If this speed is greater than the ego's speed, the new vehicle will appear at the back of the road section; if it is less, it will appear at the front. For each lane, g^* is calculated, either for the new vehicle if the appearance is at the back or for the nearest following vehicle if the appearance is at the front. The new vehicle appears in the lane where the clearance to the nearest car is greatest. If no clearance is greater than g^* , the new vehicle does not appear.

Once the ego reaches the target lane ($y = y_{\text{target}}$) or passes the distance limit ($x \geq L$), the problem terminates.

For convenience, throughout this chapter, the behavior described so far will be denoted compactly by the state transition function

$$s' = F(s, u, w). \quad (3.5)$$

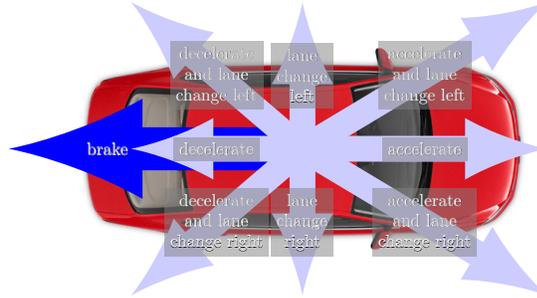


Figure 3.2: Lane changing action space

3.2.3 Action Space for Crash-Free Driving

At each time step, the planner for the ego must choose the longitudinal and lateral acceleration. For simplicity, the vehicle chooses from up to ten discrete actions which are shown in Fig. 3.2. The vehicle may make an incremental decrease or increase in speed or maintain speed, and it may begin a left or right lane change or maintain the current lane. The combination of these adjustments make up nine of the actions. The final action is a braking action determined dynamically based on the speed and position of the vehicle ahead. In most cases, the acceleration for this action is a nominal value, $-b_{\text{nominal}}$, but this is sometimes overridden. At each time step, the maximum permitted acceleration, a_{max} , is the maximum acceleration that the ego could take such that, if the vehicle ahead immediately begins braking at the physical limit, b_{max} , to a stop, the ego will still be able to stop before hitting it without exceeding physical braking limits itself. The braking action is $(\ddot{x}_e, \ddot{y}_e) = (\min\{a_{\text{max}}, -b_{\text{nominal}}\}, 0)$.

The inclusion of the dynamic braking action guarantees that there will always be an action available to the ego to avoid a crash. At each step, the action space is pruned so that if $\ddot{x}_e > a_{\text{max}}$ or if a lane change leads to a crash, that action is not considered. Since the IDM and MOBIL models are both crash-free [51], and actions that lead to crashes for the ego are not considered, no crashes occur in the simulation. Eliminating crashes in our model is justifiable because it is likely that in an actual autonomous vehicle a high-level planning system would be augmented with a low-level crash prevention system to increase safety and facilitate certification. In

addition, it is difficult to model driver behavior in the extraordinary case of a crash.

3.2.4 Reward Function and Objectives

The qualitative objectives in solving this problem are to reach the target lane within a specified distance, L , and maintain the comfort and safety of both the ego and the other nearby vehicles. Thus, the following two metrics will be used to evaluate planning performance: 1) the fraction of episodes in which the ego reaches the target lane, and 2) the fraction of episodes in which *any* vehicle operates in an unsafe manner. For this work, hard braking and unusually slow velocity are considered unsafe. A hard braking maneuver is defined as $\ddot{x} < -b_{\text{hard}}$ and slow velocity as $\dot{x} < \dot{x}_{\text{slow}}$, where b_{hard} and \dot{x}_{slow} are chosen to be uncomfortably abrupt deceleration or slow travel that might result in an accident in real conditions (see Table 3.2). In addition to quantifying safety, hard braking also serves as a proxy for comfort.

In order to encourage the planner to choose actions that will maximize these metrics, the reward function for the POMDP is defined as follows:

$$\mathcal{R}(s, a, s') = \text{in_goal}(s') - \lambda (\text{any_hard_brakes}(s, s') + \text{any_too_slow}(s')) \quad (3.6)$$

where

$$\text{in_goal}(s') = \mathbf{1}(y_e = y_{\text{target}}, x_e \leq L), \quad (3.7)$$

$$\text{any_hard_brakes}(s, s') = \max_{i \in 1..N} \{\mathbf{1}(\dot{x}'_i - \dot{x}_i < -b_{\text{hard}} \Delta t)\}, \quad (3.8)$$

$$\text{any_too_slow}(s') = \max_{i \in 1..N} \{\mathbf{1}(\dot{x}_i < \dot{x}_{\text{slow}})\}. \quad (3.9)$$

That is, there is a positive reward for reaching the target lane within the distance limit, and hard brakes and slow velocity for any car are penalized. The weight λ balances the competing goals and can be adjusted to create an approximate curve of Pareto-optimal solutions.

3.2.5 Initial Scenes

Initial scenes for the simulations are generated by beginning a simulation with only the ego on the road section and then simulating 200 steps to allow other vehicles to accumulate in the scene.

3.3 Solution Approaches

Each of the solution techniques is based on MCTS-DPW (Section 1.2.6), but handles uncertainty in the internal states differently.

3.3.1 Approach 1: Assume normal behavior

The first performance baseline is established by planning as if all cars behave according to a single static “normal” internal state (see Table 3.1). In this case, the problem is an MDP, which is solved using the MCTS-DPW algorithm. This is an overconfident baseline — it plans assuming it knows more about the other drivers than is justified by the information it has collected.

3.3.2 Approach 2: Model all uncertainty as outcome uncertainty (Naive MDP)

The second performance baseline is established by planning as if all uncertainty is simply outcome uncertainty, that is, as if the problem were an MDP with a state consisting only of the physical state and the internal states random variables, independent at each timestep, distributed according to the internal state distribution Θ . This model would be the result of fitting a Markov model with only the physical state based on data from all drivers. The MDP is again solved using the MCTS-DPW algorithm. This is a conservative baseline — it plans assuming it can learn nothing new about the drivers.

3.3.3 Approach 3: Mean Model Predictive Control

Since information about a human’s internal state can be inferred by observing a car’s physical motion, performance superior to either of the baselines can be achieved by estimating θ online. This is accomplished with a particle filter (Section 1.2.7). Filtering is independent for each car, but all of the behavior parameters for a given car are estimated jointly. There are two versions of the filter. In the first version, a particle, $\hat{\theta}$, consists of values of all model parameters. In the second version, all parameters are assumed perfectly correlated (see Section 3.4.1), so a particle consists of only a single value, the “aggressiveness”.

The belief at a given time consists of the exactly known physical state, q , and a collection of M particles, $\{\hat{\theta}^k\}_{k=1}^M$, along with associated weights, $\{W^k\}_{k=1}^M$. To update the belief when action u is taken, M new particles are sampled with probability proportional to the weights, and sampled noise values $\{\hat{w}^k\}_{k=1}^M$ are used to generate new states according to $\hat{s}^{k'} = F((q, \hat{\theta}^k), u, \hat{w}^k)$. The new weights are determined by approximating the conditional probability of the particle given the observation:

$$W^{k'} = \left\{ \begin{array}{ll} \max \left\{ 0, \frac{a-2|\dot{x}' - \hat{\dot{x}}'|}{a} \right\} & \text{if } y' = \hat{y}' \\ \gamma_{\text{lane}} \max \left\{ 0, \frac{a-2|\dot{x}' - \hat{\dot{x}}'|}{a} \right\} & \text{o.w.} \end{array} \right\} \propto \Pr(\hat{\theta}^k | o)$$

where \dot{x}' and y' are taken from the observation, $\hat{\dot{x}}'$ and \hat{y}' are from $\hat{s}^{k'}$, the max expression is proportional to the probability density of the acceleration noise triangular distribution, and $\gamma_{\text{lane}} \in [0, 1]$ is a hand-tuned parameter that penalizes incorrect lane changes (see Table 3.2).

Model predictive control (MPC) is a widely used family of control techniques that use an imperfect model and feedback measurements to choose actions [34]. At each time step, a model predictive controller calculates a sequence of control actions that will maximize a reward function of the states visited up to a future horizon given that the system behaves according to a model. The first control action in this optimized sequence is executed, and the process is repeated after a new measurement is received.

In the mean model predictive control (MMPC) approach, this particle filter is

used to estimate the internal state for each driver. At each step, MPC uses the MDP that results from assuming that each driver has the internal state corresponding to the mean of the particles in the belief approximation as the model for planning. Each time a new observation is received, the particle filter is updated and MCTS-DPW determines the best action for the resulting MDP.

3.3.4 Approach 4: QMDP

The fourth approach uses MCTS-DPW to solve the QMDP approximation of the POMDP (see Section 1.2.9).

3.3.5 Approach 5: POMCPOW

The final approach uses the POMCPOW solver described in Section 4.2.3. Though there is no theoretical guarantee that this approach will converge to an optimality, this is the closest approximation to the exact POMDP solution.

3.4 Results

The computational results from this study are designed to meet the two goals of 1) quantifying the size of the gap between the baseline control algorithm and the maximum potential lane change performance and 2) showing which cases internal state estimation and POMDP planning can approach the upper bound on performance. Experiments are carried out in three scenarios, each with a different distribution of internal states. In each of these scenarios, each of the approaches described in Section 3.3 are compared with an approximate upper performance bound obtained by planning with perfect knowledge of the behavior models.

3.4.1 Driver Model Distribution Scenarios

For the numerical testing, three internal state distribution scenarios were considered. In all of these scenarios, drivers behave according to the IDM and MOBIL models presented in Section 3.2.1, however the IDM and MOBIL parameter values are distributed differently.

Table 3.1 shows typical parameter values for aggressive, timid, and normal drivers. The values are taken from Kesting, Treiber, and Helbing [51], but some have been adjusted slightly so that the parameters for the normal driver are exactly half way between values for the timid and aggressive drivers. In all three of the scenarios, the *marginal* distributions of the parameters are uniformly distributed between the aggressive and timid values.

The difference between the scenarios is the correlation of the parameter values. In order to maintain consistent marginal distributions while varying the correlation, copulas are used. A copula is a multivariate distribution with uniform marginal distributions, but a joint distribution that causes the variables to have nonzero correlation [74]. An n -dimensional Gaussian copula is defined by a correlation matrix, Σ , and has a cumulative distribution function defined as follows:

$$F_{GC}(x) = \Phi_{\Sigma}(\Phi^{-1}(x_1), \dots, \Phi^{-1}(x_n)), \quad (3.10)$$

where Φ_{Σ} is the cumulative distribution function for a multivariate Gaussian distribution with covariance Σ , and Φ^{-1} is the inverse cumulative distribution function for a univariate Gaussian distribution. If random vector X has cumulative distribution function F_{GC} , then for any $i, j \in 1..n, i \neq j$, the correlation between X_i and X_j is Σ_{ij} .

In Scenario 1, all of the parameters are independently distributed. In Scenario 2, all of the parameters are perfectly correlated so that all parameters are deterministic functions of a single uniformly-distributed random variable, the “aggressiveness” of the driver. In Scenario 3, the distribution is correlated between these two extremes.

Specifically, it is a Gaussian copula with covariance matrix

$$\Sigma = \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & & \\ \vdots & & \ddots & \vdots \\ & & & 1 & \rho \\ \rho & \dots & \rho & \rho & 1 \end{bmatrix}, \quad (3.11)$$

that is, a matrix with 1 along the diagonal and correlation parameter ρ elsewhere. The values drawn from this distribution are scaled and translated to lie between the aggressive and normal limits.

For Scenario 3, the value of ρ is 0.75, and Scenarios 1 and 2 can be thought of as limiting cases where ρ approaches 0 and 1, respectively. In Scenarios 1 and 3, the first version of the particle filter, which estimates all of the model parameters jointly, is used, whereas in Scenario 2, the second version of the particle filter that assumes fully correlated parameters is used, that is, it only estimates a single “aggressiveness” parameter for each car. The mean MPC planner uses this “aggressiveness” parameter for all scenarios because this resulted in better performance. The small scatter plots in Figs. 3.3 and 3.4 illustrate the level of correlation by plotting sampled values of two of the parameters.

Table 3.1: IDM and MOBIL parameters for different driver types.

IDM Parameter		Timid	Normal	Aggressive
Desired speed (m/s)	\dot{x}_0	27.8	33.3	38.9
Desired time gap (s)	T	2.0	1.5	1.0
Jam distance (m)	g_0	4.0	2.0	0.0
Max acceleration (m/s ²)	a	0.8	1.4	2.0
Desired deceleration (m/s ²)	b	1.0	2.0	3.0
MOBIL Parameter		Timid	Normal	Aggressive
Politeness	p	1.0	0.5	0.0
Safe braking (m/s ²)	b_{safe}	1.0	2.0	3.0
Acceleration threshold (m/s ²)	a_{thr}	0.2	0.1	0.0

Table 3.2: Various simulation parameters

Parameter	Symbol	Value
Simulation time step	Δt	0.75 s
Max vehicles on road	N_{\max}	10
Lane change rate	\dot{y}_{lc}	0.67 lanes/s
Distance limit	L	1000 m
Velocity noise standard deviation	σ_{vel}	0.5 m/s
Physical braking limit	b_{\max}	8.0 m/s ²
Penalized hard braking limit	b_{hard}	4.0 m/s ²
Penalized minimum speed	\dot{x}_{slow}	15 m/s
UCT exploration parameter	c	8
DPW linear parameter	k	4.5
DPW exponent parameter	α	0.1
MCTS search depth		40
MCTS iterations per step		1000
Particle filter wrong lane factor	γ_{lane}	0.05
Number of Particles (Joint Parameter Filter)	M	5000
Number of Particles (Aggressiveness Filter)	M	2000
Reward ratios for points on approximate Pareto curves	λ	0.5, 1, 2, 4, 8

3.4.2 Pareto Front Comparison

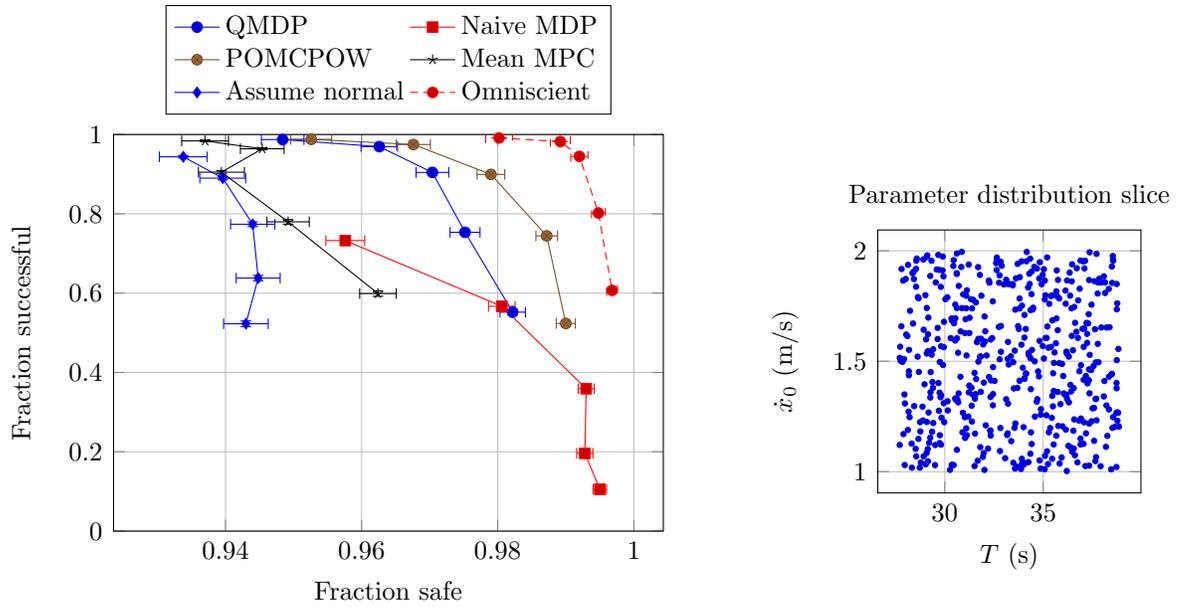
Figures 3.3 and 3.4 show approximate Pareto fronts illustrating the performance in terms of safety and efficiency of each of the approaches described in Section 3.3. Each of the points on the curve shows the results of 5000 independent simulations of the scenario with a particular safety-efficiency tradeoff weight, λ .

The baseline and upper bound approaches perform as expected. The baseline planner that assumes all vehicles act with normal behavior parameters creates overconfident plans. That is, it is able to reach the goal a large proportion of the time, but it causes many safety violations. On the other hand, the naive MDP approximation is over-cautious. That is, it can attain a high level of safety, but it is never able to meet the goal more than 80% of the time. The omniscient upper bound planner achieves performance equal to or greater than all other approaches.

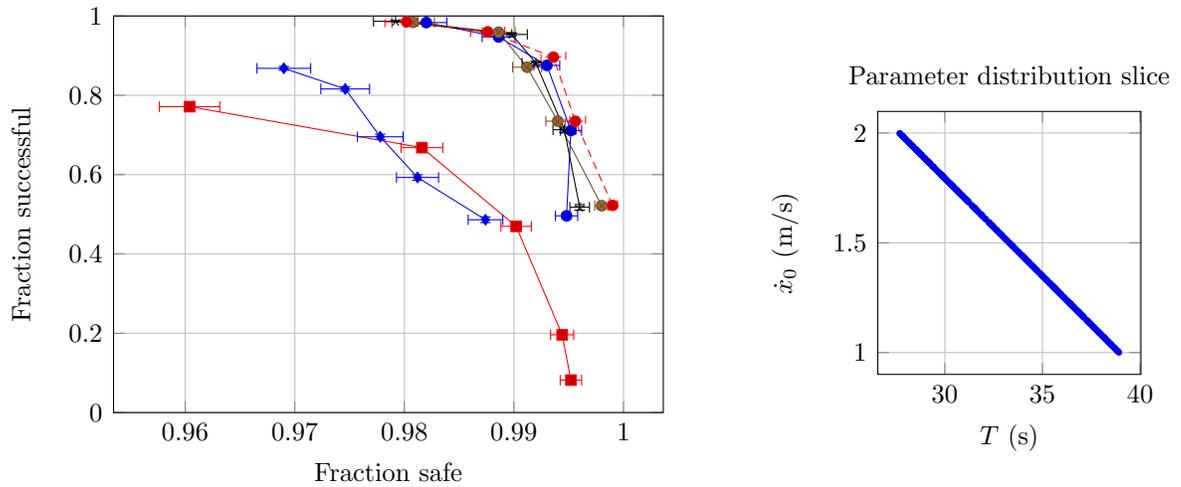
As expected, better plans are attained as more accurate uncertainty is modeled in planning. The mean MPC approach usually performs better than the baselines because it dynamically estimates the model parameters, but it is still overconfident (achieving a high success rate, but sacrificing safety) because it plans without any internal state uncertainty. QMDP performs better than mean MPC because it considers samples from the entire estimated internal state distribution when planning. Since the vehicle does not have to take costly information-gathering actions to accomplish its goal, POMCPOW only rarely outperforms QMDP.

Convexity violations

One immediate concern that should be raised about the approximate Pareto frontiers in Figs. 3.3 and 3.4 is that they are not all convex. The Pareto-optimal points generated by solving optimization problems of the form in (1.4) must lie on the convex hull of the true Pareto front (see Boyd and Vandenberghe [14], Example 2.27). Thus, approximate Pareto fronts plotted by connecting particular solutions with straight lines as in Figs. 3.3 and 3.4 should be convex. Particularly egregious violations of convexity can be found in the mean MPC curve in Fig. 3.3a and the normal behavior assumption curve in Fig. 3.4, where there are “kinks” at the third point from the top



(a) Scenario 1: Uncorrelated ($\rho = 0$)



(b) Scenario 2: Fully correlated ($\rho = 1$)

Figure 3.3: Approximate Pareto performance curves for model distributions at correlation extremes. The scatter plots at right illustrate the level of correlation with samples from the joint parameter distribution. Error bars indicate the standard error of the mean.

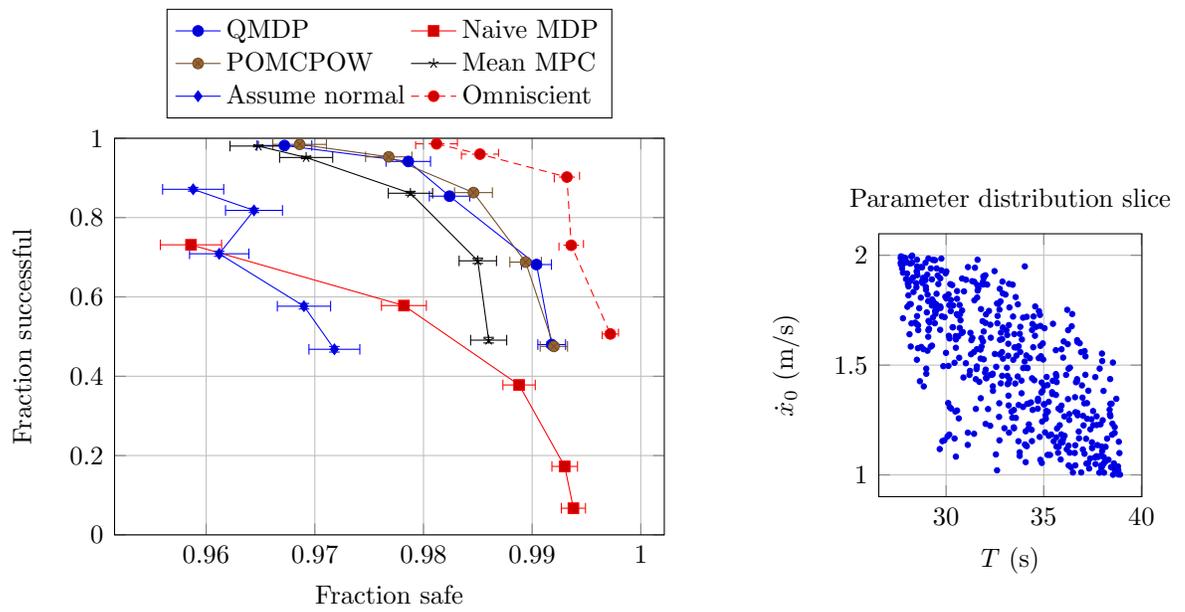


Figure 3.4: Approximate Pareto performance curves for partially correlated model distribution (Scenario 3: $\rho = 0.75$). Error bars indicate the standard error of the mean.

($\lambda = 2$) that prevent these curves from even being monotonic.

The lack of convexity may be due to some combination of the following reasons:

1. The performance objectives plotted in the graphs do not exactly match the stage-wise reward function (3.6). For example, the planner observes a larger penalty if there are multiple safety violations, but this is not reflected in the plots.
2. The MCTS-DPW solution method is itself stochastic and has no guarantees of convergence in finite time.
3. Even given infinite computing time, the solvers will converge to inaccurate approximations of the true POMDP solution (except, perhaps, for POMCPOW).

One compelling explanation for the kinks mentioned above is that, as λ is increased, since the planner is penalized more severely for unsafe actions, it plans a more conservative trajectory and stays on the road longer. The longer time on the road gives more chances for unsafe events to occur which are difficult for the planner to predict and avoid because of its inaccurate model. This explanation is corroborated by the results in Fig. 3.5. In both places where there were previously kinks, the number of hard brakes per kilometer decreases as λ increases.

3.4.3 Correlation comparison

It is also interesting to consider the effect that the correlation between model parameters has on the relative effectiveness of the control approaches. Figure 3.3a shows that when there is no correlation, QMDP offers a significant advantage over mean MPC, and POMCPOW offers a further significant advantage over QMDP. In this case, since the parameters are uncorrelated, there is a large amount of uncertainty in them even when some (e.g. \dot{x}_0) are easy to observe, and since POMCPOW is able to plan into the future considering this uncertainty, it performs better. On the other hand, when the parameters are fully correlated as shown in Fig. 3.3b, all of the parameters are easy to estimate by observing only a few, so there is not a significant performance gap between mean MPC, QMDP, and POMCPOW; all are able to close

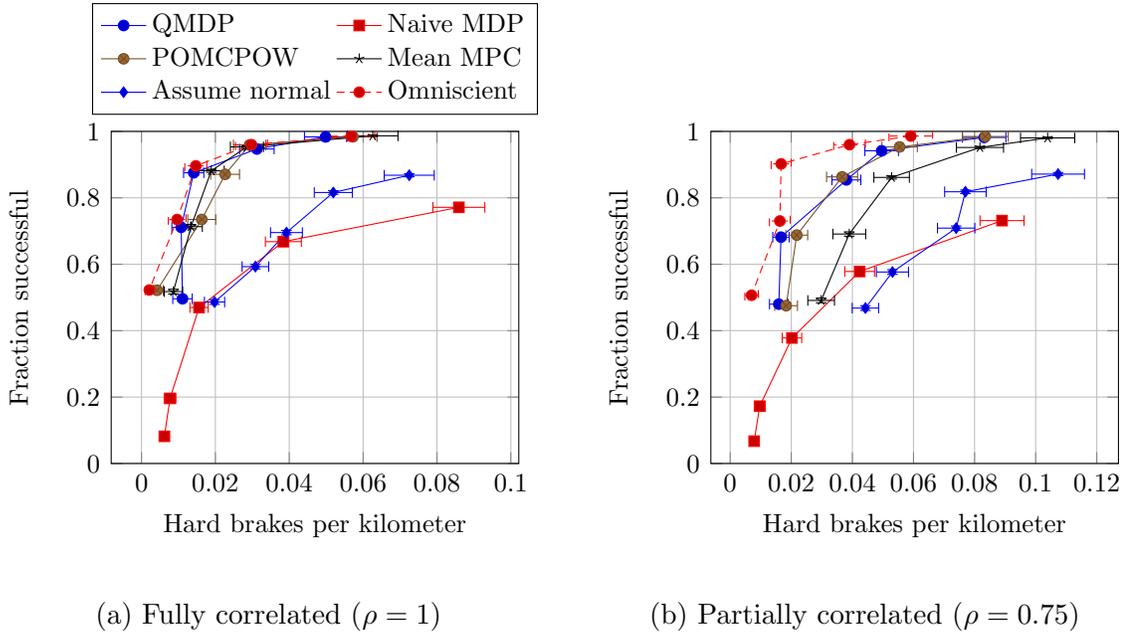


Figure 3.5: Average hard braking frequency and success rate

the gap and achieve nearly the same performance as the upper bound. Figure 3.4 shows the expected behavior between the extremes.

Figure 3.6 shows the performance gaps at more points between $\rho = 0$ and 1. As the correlation increases, the approximate POMDP planning approaches get steadily closer to closing the performance gap with the upper bound. These results have significant implications for the real world. It suggests that if most human driver behavior is correlated with easily measurable quantities, near-optimal performance can be achieved by simpler approaches like mean MPC. If there is little correlation, planning with internal states offers a less pronounced benefit, and more advanced planners that carry the uncertainty further into the future are needed to realize that benefit.

3.4.4 Robustness

In the experiments above, internal parameter distribution is assumed to be known exactly. Because there will be differences between any model used in planning and

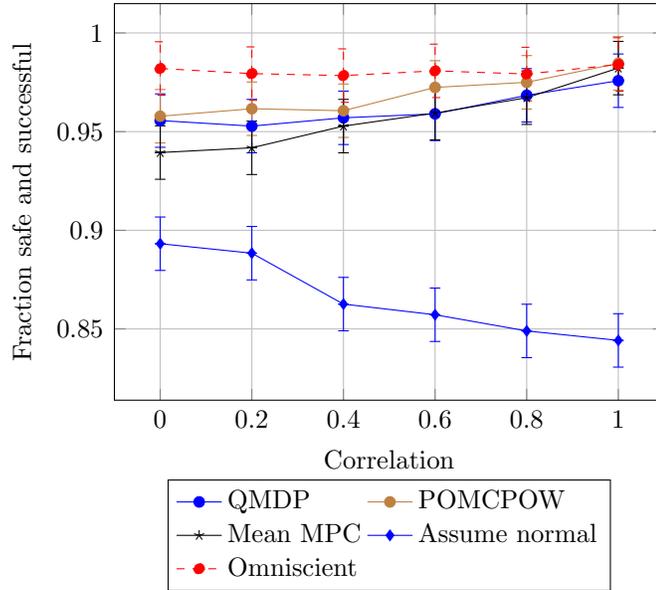


Figure 3.6: Performance variation with Θ correlation. Error bars indicate the 68% (corresponding to one standard deviation in a normal distribution) confidence region determined by the Hoeffding Bound. The Naive MDP performance is not shown because it is significantly lower than the other approaches.

the way human drivers actually behave, it is important to test planning algorithms for robustness. This section contains tests in which the parameter distribution for planning differs from the true distribution.

Parameter Correlation Robustness

The first robustness test examines the effect of correlation inaccuracy. POMCPOW and QMDP planners that assume no correlation and full correlation are tested against simulation models with varying levels of correlation. Figure 3.7 shows the results. Performance does not degrade abruptly when the correlation model is inaccurate, though there is an advantage to planning with a correlated model when the true parameters are fully correlated.

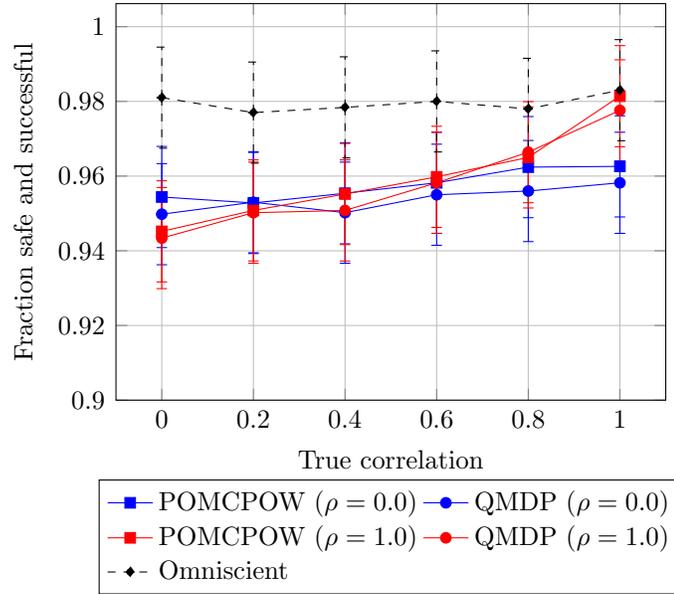


Figure 3.7: Parameter correlation robustness study. Error bars indicate the 68% confidence region determined by the Hoeffding Bound.

Parameter Domain Robustness

In the second robustness test, the domain from which the parameters are drawn is expanded by a variable factor. Specifically, in the new test distribution, the “Normal” values from Table 3.2 are still used as the midpoints of the distribution, but the distance to the extremes is multiplied by the expansion factor. For example, if the expansion factor is 2, the marginal distribution of the desired time gap, T , is uniform between 0.5 s and 2.5 s. Distributions that contain physically nonsensical values such as a negative jam distance, g_0 , are truncated. The planner always uses the distribution defined by values from Table 3.2.

Figure 3.8 shows the results of tests with expansion factor values between 0.2 and 2.0. All approaches have significantly more success when the expansion factor is less than 1, indicating that the problem is easier in this case, even with an inaccurate model. There is significant performance degradation at higher expansion factors, both in absolute terms and relative to the omniscient upper bound. However, this degradation is one-sided; there is little performance lost due to planning with distribution

domains that are larger than the true distribution domain. This one-sidedness suggests that, in practice, prior distributions should be chosen conservatively (i.e. with a larger domain than that of the true distribution) so that the planner can handle situations with both higher- and lower-than-expected levels of uncertainty.

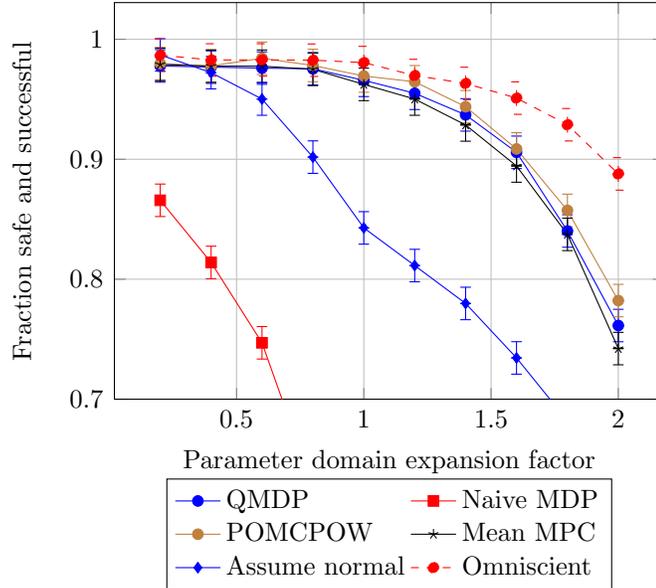


Figure 3.8: Parameter domain robustness study. Error bars indicate the 68% confidence region determined by the Hoeffding Bound.

3.5 Discussion

This chapter investigates the effects of modeling outcome and state uncertainty in autonomous driving. The problem of making multiple lane changes on a highway within a limited distance is modeled as a POMDP. Both MDP solutions that ignore the internal states of other drivers, and POMDP solutions that do model these internal states with various degrees of approximation are considered. An upper performance bound is established by planning with full knowledge of the internal states. In order to remain agnostic with regard to the relative importance of safety and efficiency, Pareto frontiers are compared.

POMDP approaches are compared to both a conservative MDP formulation that

models internal state uncertainty as outcome uncertainty and an overly confident MDP approach where knowledge of the internal state is assumed. The advantage of the POMDP approaches is clear in all of the tests. However, the relative effectiveness of different POMDP approaches is heavily dependent on the correlation of the distributions of the internal states. If the internal states are highly correlated, simply estimating them with a particle filter and planning assuming certainty equivalence is adequate to nearly match the upper performance bound. On the other hand, when the parameters are uncorrelated, the QMDP planner performs much better than the certainty equivalence planner, and POMCPOW performs much better than QMDP. Moreover, in this uncorrelated case, there is a significant gap between all approaches and the upper bound. Partial correlation produces the expected results between the correlated and uncorrelated cases.

Experiments also characterize the robustness of the algorithms to incorrect parameter distributions. POMCPOW and QMDP do not suffer significant performance degradation when the parameter correlation is not correct. Robustness to inaccuracy in the parameter domain is one-sided: when the true domain is larger than that assumed by the planners, performance is adversely affected, but when the true domain is smaller, there is no degradation.

The primary weakness of this investigation is the model for other drivers. Since the IDM and MOBIL models were developed to simulate large scale traffic flow [106, 50], their accuracy for microscopic simulations has not been attested with data. Upgrading to models learned from data would further validate the conclusions drawn here. Other similar research has taken steps in this direction by using a conditional variational auto-encoder model [88]. The model used here also neglects one of the key internal states of drivers: intentions. Planning based on the possible intentions of other drivers would likely be an even more powerful planning technique than the approach investigated here because it would enable sophisticated interaction and communication between the autonomous cars and humans.

The results of this investigation should also be used to guide future research into modeling. In particular, since correlation is such a significant factor influencing performance, it is important to determine whether internal states that determine

behavior are correlated with easily measured quantities in real human drivers.

Finally, more advanced planning techniques, such as the DESPOT algorithm, should also be investigated. In order to isolate the effects of uncertainty modeling, the solution algorithms in this investigation were limited to variants of Monte Carlo tree search that use double progressive widening. The next chapter analyzes online algorithms for solving POMDPs with continuous observation spaces like the problem addressed here. In addition to proposing the POMCPOW algorithm used in this chapter, it includes a comparison with DESPOT on this problem and shows that it outperforms POMCPOW because of its ability to plan with a longer effective horizon.

Chapter 4

Online Algorithms for Continuous POMDPs

The autonomous driving problem in the previous chapter illustrates that modeling partially observable internal states can have a significant effect on performance. POMDPs based on real world problems, such as the one considered in Chapter 3, often have continuous state, action, and observation spaces. Although the research surveyed in the introduction (Section 1.2.8) has yielded effective solution techniques for many classes of POMDPs, there remains a need for simple, general purpose online solvers that can handle continuous spaces, especially continuous observation spaces. This chapter takes some important steps toward addressing that need.

POMCP and other online methods can easily accommodate continuous state spaces without any modification [38]. However, there has been less progress on problems with continuous observation spaces. This chapter presents two similar algorithms which address the challenge of solving POMDPs with continuous state, action, and observation spaces. The first solves an approximation of the belief-space MDP and is called particle filter trees with double progressive widening (PFT-DPW). The second is based on POMCP and is called partially observable Monte Carlo planning with observation widening (POMCPOW).

4.1 Background

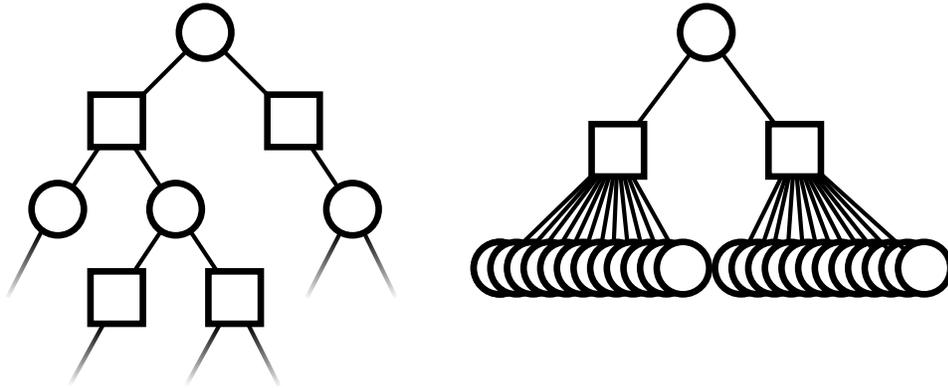


Figure 4.1: POMCP tree for a discrete POMDP (left), and for a POMDP with a continuous observation space (right). Because the observation space is continuous, each simulation creates a new observation node and the tree cannot extend deeper.

There are two challenges that make tree search difficult in continuous spaces. The first is that, since the probability of sampling the same real number twice from a continuous random variable is zero, the width of the planning trees explodes on the first step, causing them to be too shallow to be useful (see Fig. 4.1). POMCPOW and PFT-DPW resolve this issue with a technique called double progressive widening (DPW) [23]. The second issue is that, even when DPW is applied, the belief representations used by current solvers collapse to a single state particle, resulting in overconfidence. As a consequence, the solutions obtained resemble QMDP policies, and there is no incentive for information gathering. POMCPOW and PFT-DPW overcome this issue by using the observation model to weight the particles used to represent beliefs.

A small amount of previous research has sought online solutions to continuous POMDPs. ABT has been extended to use generalized pattern search for selecting locally optimal continuous actions, an approach which is especially effective in problems where high precision is important [90, 109]. Continuous observation Monte Carlo tree search (COMCTS) constructs observation classification trees to automatically partition the observation space in a POMCP-like approach; however it did not perform

much better than a Monte Carlo rollout approach in experiments [77].

4.2 Algorithms

This section presents and discusses several methods for handling continuous POMDPs. First, it presents three new MCTS-DPW-based algorithms. The first of these is POMCP-DPW. Theorem 1 shows that this algorithm is suboptimal in some cases. The second algorithm is PFT-DPW, a straightforward application of MCTS-DPW to the belief MDP with particle filters to approximate belief updates. The third algorithm is POMCPOW, which fixes the problems noted about POMCP-DPW and successfully extends POMCP to problems with continuous observation spaces. These three algorithms are followed by brief discussion about the discretization alternative to these algorithms and the additional observation distribution requirement of POMCPOW and PFT-DPW.

The three new algorithms in this chapter share some common structure. For all algorithms, the entry point for the decision making process is the `PLAN` procedure, which takes the current belief, b , as an input (`PLAN` differs slightly for PFT-DPW in Algorithm 4). The algorithms also share the same `ACTIONPROGWIDEN` function to control progressive widening of the action space. These components are listed in Algorithm 2. The difference between the algorithms is in the `SIMULATE` function.

The following variables are used in the listings and text: h represents a history $(b, a_1, o_1, \dots, a_k, o_k)$, and ha and hao are shorthand for histories with a and (a, o) appended to the end, respectively; d is the depth left to explore, with d_{\max} the maximum depth; C is a list of the children of a node (along with the reward in the case of PFT-DPW); N is a count of the number of visits; and M is a count of the number of times that a history has been generated by the model. The list of states associated with a node is denoted B , and W is a list of weights corresponding to those states. Finally, $Q(ha)$ is an estimate of the value of taking action a after observing history h . C , N , M , B , W , and Q are all implicitly initialized to 0 or \emptyset . The `ROLLOUT` procedure runs a d -step simulation with a default rollout policy, which can be based on the history or fully observed state, and returns the discounted reward.

Algorithm 2 Common procedures

```

1: procedure PLAN( $b$ )
2:   for  $i \in 1 : n$ 
3:      $s \leftarrow$  sample from  $b$ 
4:     SIMULATE( $s, b, d_{\max}$ )
5:   return  $\operatorname{argmax}_a Q(ba)$ 

6: procedure ACTIONPROGWIDEN( $h$ )
7:   if  $|C(h)| \leq k_a N(h)^{\alpha_a}$ 
8:      $a \leftarrow$  NEXTACTION( $h$ )
9:      $C(h) \leftarrow C(h) \cup \{a\}$ 
10:  return  $\operatorname{argmax}_{a \in C(h)} Q(ha) + c \sqrt{\frac{\log N(h)}{N(ha)}}$ 

```

4.2.1 POMCP-DPW

The first algorithm that we consider is POMCP with double progressive widening (POMCP-DPW). In this algorithm, listed in Algorithm 3, the number of new children sampled from any node in the tree is limited by DPW using the parameters k_a , α_a , k_o , and α_o . In the case where the simulated observation is rejected (line 14), the tree search is continued with an observation selected in proportion to the number of times, M , it has been previously simulated (line 15), and a state is sampled from the associated belief (line 16).

This algorithm may converge to the optimal solution for POMDPs with discrete observation spaces; however, on continuous observation spaces, POMCP-DPW is suboptimal. In particular, it finds a QMDP policy (see Section 1.2.9). In fact, for a modified version of POMCP-DPW, it is easy to prove that it will converge to such a policy (Theorem 1, below). Appendix A describes the modified algorithm and problem requirements including the definitions of polynomial exploration, the regularity hypothesis for the problem, and exponentially sure convergence that are used in the discussion of the theorem below.

Definition 1 (QMDP value). *Let $Q_{MDP}(s, a)$ be the optimal state-action value function assuming full observability starting by taking action a in state s . The QMDP value at belief b , $Q_{MDP}(b, a)$, is the expected value of $Q_{MDP}(s, a)$ when s is distributed*

Algorithm 3 POMCP-DPW

```

1: procedure SIMULATE( $s, h, d$ )
2:   if  $d = 0$ 
3:     return 0
4:    $a \leftarrow \text{ACTIONPROGWIDEN}(h)$ 
5:   if  $|C(ha)| \leq k_o N(ha)^{\alpha_o}$ 
6:      $s', o, r \leftarrow G(s, a)$ 
7:      $C(ha) \leftarrow C(ha) \cup \{o\}$ 
8:      $M(hao) \leftarrow M(hao) + 1$ 
9:      $B(hao) \leftarrow B(hao) \cup \{s'\}$ 
10:    if  $M(hao) = 1$ 
11:       $total \leftarrow r + \gamma \text{ROLLOUT}(s', hao, d - 1)$ 
12:    else
13:       $total \leftarrow r + \gamma \text{SIMULATE}(s', hao, d - 1)$ 
14:    else
15:       $o \leftarrow \text{select } o \in C(ha) \text{ w.p. } \frac{M(hao)}{\sum_o M(hao)}$ 
16:       $s' \leftarrow \text{select } s' \in B(hao) \text{ w.p. } \frac{1}{|B(hao)|}$ 
17:       $r \leftarrow \mathcal{R}(s, a, s')$ 
18:       $total \leftarrow r + \gamma \text{SIMULATE}(s', hao, d - 1)$ 
19:     $N(h) \leftarrow N(h) + 1$ 
20:     $N(ha) \leftarrow N(ha) + 1$ 
21:     $Q(ha) \leftarrow Q(ha) + \frac{total - Q(ha)}{N(ha)}$ 
22:    return  $total$ 

```

according to b .

Theorem 1 (Modified POMCP-DPW convergence to QMDP). *If a bounded-horizon POMDP meets the following conditions: 1) the state and observation spaces are continuous with a finite observation probability density function, and 2) the regularity hypothesis is met, then modified POMCP-DPW will produce a value function estimate, \hat{Q} , that converges to the QMDP value for the problem. Specifically, there exists a constant $C > 0$, such that after n iterations,*

$$\left| \hat{Q}(b, a) - Q_{MDP}(b, a) \right| \leq \frac{C}{n^{1/(10d_{\max}-7)}}$$

exponentially surely in n , for every action a .

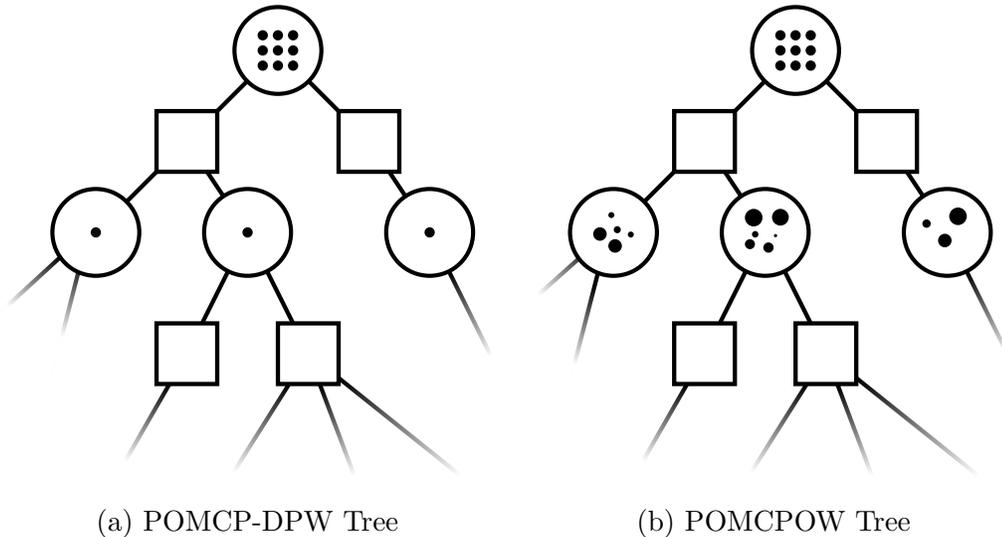


Figure 4.2: Tree structure comparison. Each square is an action node, and each unfilled circle is an observation node. Each black dot corresponds to a state particle with the size representing its weight. In continuous observation spaces, the beliefs in a POMCP-DPW tree degenerate to a single particle, while POMCPOW maintains weighted particle mixture beliefs.

A proof of this theorem that leverages work by Auger, Couetoux, and Teytaud [4] is given in Appendix A, but a brief justification is provided here. The key is that belief nodes will contain only a single state particle (see Fig. 4.2). This is because, since the

observation space is continuous with a finite density function, the generative model will (with probability one) produce a unique observation o each time it is queried. Thus, for every generated history h , only one state will ever be inserted into $B(h)$ (Algorithm 3, line 9), and therefore h is merely an alias for that state. Since each belief node corresponds to a state, the solver is actually solving the fully observable MDP at every node except the root node, leading to a QMDP solution.

As a result of Theorem 1, the action chosen by modified POMCP-DPW will match a QMDP policy (a policy of actions that maximize the QMDP value) with high precision exponentially surely (see Corollary 1 of Auger, Couetoux, and Teytaud [4]). For many problems, e.g. the problem in Chapter 3, this is a very useful solution, but since it neglects the value of information, a QMDP policy is suboptimal for problems where information gathering is important [65, 52].

Although Theorem 1 is only theoretically applicable to the modified version of POMCP-DPW, it helps explain the behavior of other solvers. Modified POMCP-DPW, POMCP-DPW, DESPOT, and ABT all share the characteristic that a belief node can only contain two states if they generated exactly the same observation. Since this is an event with zero probability for a continuous observation space, these solvers exhibit suboptimal, often QMDP-like, behavior. The experiments in Section 4.3.6 show this for POMCP-DPW and DESPOT, and this is presumably the case for ABT as well.

4.2.2 PFT-DPW

Another algorithm that one might consider for solving continuous POMDPs online is MCTS-DPW on the equivalent belief MDP. Since the Bayesian belief update is usually computationally intractable, a particle filter is used. This new approach will be referred to as particle filter trees with double progressive widening (PFT-DPW). It is shown in Algorithm 4, where $G_{\text{PF}(m)}(b, a)$ is a particle filter belief update performed with a simulated observation and m state particles which approximates the belief MDP generative model. The authors are not aware of any mention of this algorithm in prior literature, but it is very likely that MCTS with particle filters has been used

Algorithm 4 PFT-DPW

```

1: procedure PLAN( $b$ )
2:   for  $i \in 1 : n$ 
3:     SIMULATE( $b, d_{\max}$ )
4:   return  $\operatorname{argmax}_a Q(ba)$ 

5: procedure SIMULATE( $b, d$ )
6:   if  $d = 0$ 
7:     return 0
8:    $a \leftarrow \text{ACTIONPROGWIDEN}(b)$ 
9:   if  $|C(ba)| \leq k_o N(ba)^{\alpha_o}$ 
10:     $b', r \leftarrow G_{\text{PF}(m)}(b, a)$ 
11:     $C(ba) \leftarrow C(ba) \cup \{(b', r)\}$ 
12:     $total \leftarrow r + \gamma \text{ROLLOUT}(b', d - 1)$ 
13:  else
14:     $b', r \leftarrow \text{sample uniformly from } C(ba)$ 
15:     $total \leftarrow r + \gamma \text{SIMULATE}(b', d - 1)$ 
16:   $N(b) \leftarrow N(b) + 1$ 
17:   $N(ba) \leftarrow N(ba) + 1$ 
18:   $Q(ba) \leftarrow Q(ba) + \frac{total - Q(ba)}{N(ba)}$ 
19:  return  $total$ 

```

before without double progressive widening under another name.

PFT-DPW is fundamentally different from POMCP and POMCPOW because it relies on simulating approximate belief trajectories instead of state trajectories. This distinction also allows it to be applied to problems where the reward is a function of the belief rather than the state such as pure information-gathering problems [28, 3].

The primary shortcoming of this algorithm is that the number of particles in the filter, m , must be chosen a-priori and is static throughout the tree. Each time a new belief node is created, an $\mathcal{O}(m)$ particle filter update is performed. If m is too small, the beliefs may miss important states, but if m is too large, constructing the tree is expensive. Fortunately, the experiments in Section 4.3 show that it is often easy to choose m in practice; for all the problems studied here, a value of $m = 20$ resulted in good performance.

4.2.3 POMCPOW

In order to address the suboptimality of POMCP-DPW, a new algorithm, POMCPOW, shown in Algorithm 5 is proposed here. In this algorithm, the belief updates are weighted, but they also expand gradually as more simulations are added. Furthermore, since the richness of the belief representation is related to the number of times the node is visited, beliefs that are more likely to be reached by the optimal policy have more particles. At each step, the simulated state is inserted into the weighted particle collection that represents the belief (line 10), and a new state is sampled from that belief (line 16). A simple illustration of the tree is shown in Figure 4.2 to contrast with a POMCP-DPW tree. Because the resampling in line 16 can be efficiently implemented with binary search, the computational complexity is $\mathcal{O}(nd \log(n))$.

4.2.4 Discretization

Discretization is perhaps the most straightforward way to deal with continuous observation spaces. In this approach, the continuous observation space is simply divided into small discrete regions and solved as a discrete POMDP using conventional methods. The results in Table 4.1 show that this approach is only sometimes effective.

Algorithm 5 POMCPOW

```

1: procedure SIMULATE( $s, h, d$ )
2:   if  $d = 0$ 
3:     return 0
4:    $a \leftarrow$  ACTIONPROGWIDEN( $h$ )
5:    $s', o, r \leftarrow G(s, a)$ 
6:   if  $|C(ha)| \leq k_o N(ha)^{\alpha_o}$ 
7:      $M(hao) \leftarrow M(hao) + 1$ 
8:   else
9:      $o \leftarrow$  select  $o \in C(ha)$  w.p.  $\frac{M(hao)}{\sum_o M(hao)}$ 
10:  append  $s'$  to  $B(hao)$ 
11:  append  $\mathcal{Z}(o \mid s, a, s')$  to  $W(hao)$ 
12:  if  $o \notin C(ha)$  ▷ new node
13:     $C(ha) \leftarrow C(ha) \cup \{o\}$ 
14:     $total \leftarrow r + \gamma$ ROLLOUT( $s', hao, d - 1$ )
15:  else
16:     $s' \leftarrow$  select  $B(hao)[i]$  w.p.  $\frac{W(hao)[i]}{\sum_{j=1}^m W(hao)[j]}$ 
17:     $r \leftarrow \mathcal{R}(s, a, s')$ 
18:     $total \leftarrow r + \gamma$ SIMULATE( $s', hao, d - 1$ )
19:   $N(h) \leftarrow N(h) + 1$ 
20:   $N(ha) \leftarrow N(ha) + 1$ 
21:   $Q(ha) \leftarrow Q(ha) + \frac{total - Q(ha)}{N(ha)}$ 
22:  return  $total$ 

```

4.2.5 Observation Distribution Requirement

It is important to note that, while POMCP, POMCP-DPW, and DESPOT only require a generative model of the problem, both POMCPOW and PFT-DPW require a way to query the relative likelihood of different observations (\mathcal{Z} in line 11). One possible objection against POMCPOW and PFT-DPW is that this requirement limits them to a restricted class of POMDPs. However, there are at least two reasons to expect that it will be an effective tool in practice.

First, this requirement is no more stringent than the requirement for a standard importance resampling particle filter, and such filters are used widely, at least in the field of robotics that the authors are most familiar with. Moreover, if the observation model is complex, an approximate model may be sufficient.

Second, given the implications of Theorem 1, it is difficult to imagine a tree-based decision-making algorithm or a robust belief updater that does not require some way of measuring whether a state belongs to a belief or history. The observation model is a straightforward and standard way of specifying such a measure. Finally, in practice, except for the simplest of problems, using POMCP or DESPOT to repeatedly observe and act in an environment already requires more than just a generative model. For example, the authors of the original paper describing POMCP [93] use heuristic particle reinvigoration in lieu of an observation model and importance sampling.

4.3 Experiments

Numerical simulation experiments in five domains were conducted to evaluate the performance of POMCPOW and PFT-DPW compared to other solvers. The open source code for the experiments is built on the POMDPs.jl framework (Chapter 5) and hosted at <https://github.com/zsunberg/ContinuousPOMDPTreeSearchExperiments.jl>. In all experiments, the solvers were limited to 1 second of computation time per step. Belief updates were accomplished with a particle filter independent of the planner, and no part of the tree was saved for re-use on subsequent steps. All results are presented together in Table 4.1 on Page 85. Each of the five domains is described below

along with a a brief description of how the different solvers perform. Hyperparameter values and discretization details are given in Sections 4.3.6 and 4.3.7 respectively.

4.3.1 Laser Tag

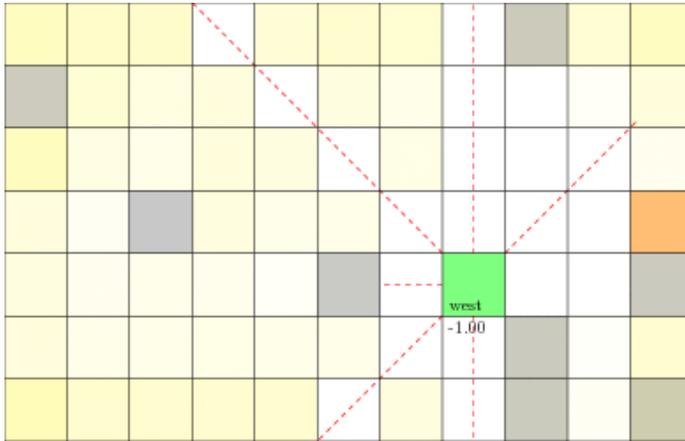


Figure 4.3: Laser Tag benchmark. The green square is the robot; the orange square is the target. The varying shades of yellow show the belief about the position of the target. The red dashed lines represent the laser sensors.

The Laser Tag benchmark (Fig. 4.3) is taken directly from the work of Somani, Ye, Hsu, and Lee [97] and included for the sake of calibration. A complete description is given in that work. DESPOT outperforms the other methods in this test. The score for DESPOT differs slightly from that reported by Somani, Ye, Hsu, and Lee [97] likely because of bounds implementation differences. POMCP performs much better than reported by Somani, Ye, Hsu, and Lee [97] because this implementation uses a state-based rollout policy.

4.3.2 Light Dark

In the Light Dark domain, the state is an integer between -60 and 60 , and the agent can choose how to move deterministically ($s' = s + a$) from the action space $\mathcal{A} = \{-10, -1, 0, 1, 10\}$. The goal is to reach the origin and take an action there. If action 0 is taken at the origin, a reward of 100 is given and the problem terminates; If

Table 4.1: Experimental results

	Laser Tag (D, D, D)	Light Dark (D, D, C)	Sub Hunt (D, D, C)
POMCPOW	-10.3 ± 0.2	56.1 ± 0.6	69.2 ± 1.3
PFT-DPW	-11.1 ± 0.2	57.2 ± 0.5	77.4 ± 1.1
QMDP	-10.5 ± 0.2	-6.4 ± 1.0	28.0 ± 1.3
POMCP-DPW	-10.6 ± 0.2	-7.3 ± 1.0	28.3 ± 1.3
DESPOT	-8.9 ± 0.2	-6.8 ± 1.0	26.8 ± 1.3
POMCP ^D	-14.1 ± 0.2	61.1 ± 0.4	28.0 ± 1.3
DESPOT ^D		54.2 ± 1.1	27.4 ± 1.3
	VDP Tag (C, C, C)	Multilane (C, D, C)	
POMCPOW	29.3 ± 0.8	30.9 ± 0.9	
PFT-DPW	27.2 ± 0.8	21.4 ± 0.9	
QMDP			
POMCP-DPW	16.4 ± 1.0	29.6 ± 0.9	
DESPOT		36.0 ± 0.8	
POMCP ^D	14.7 ± 0.9		
DESPOT ^D	14.3 ± 1.0		

The three C or D characters after the solver indicate whether the state, action, and observation spaces are continuous or discrete, respectively. For continuous problems, solvers with a superscript D were run on a version of the problem with discretized action and observation spaces, but they interacted with continuous simulations of the problem.

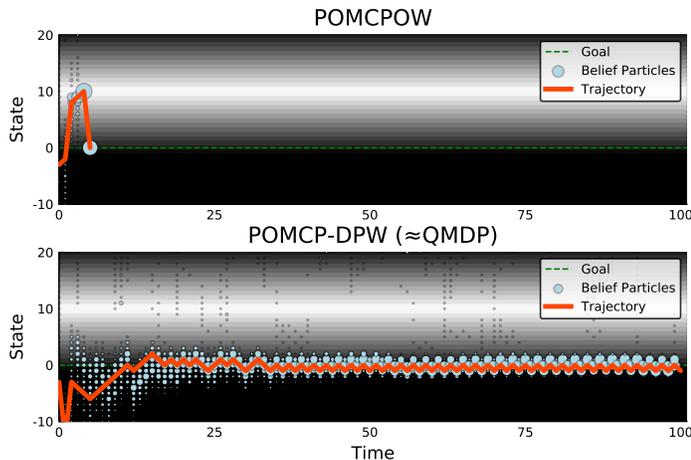


Figure 4.4: Example trajectories in the Light Dark domain. POMCPOW travels to the light region and accurately localizes before moving to the goal. POMCP-DPW displays QMDP-like behavior: It is unable to localize well enough to take action 0 with confidence. The belief particles far away from 0 in the POMCP-DPW plot are due to particle reinvigoration that makes the filter more robust.

action 0 is taken at another location, a penalty of -100 is given. There is a cost of -1 at each step before termination. The agent receives a more accurate observation in the “light” region around $s = 10$. Specifically, observations are continuous ($\mathcal{O} = \mathbb{R}$) and normally distributed with standard deviation $\sigma = |s - 10|$.

Table 4.1 shows the mean reward from 1000 simulations for each solver, and Fig. 4.4 shows an example experiment. The optimal strategy involves moving toward the light region and localizing before proceeding to the origin. QMDP and solvers predicted to behave like QMDP attempt to move directly to the origin, while POMCPOW and PFT-DPW perform better. In this one-dimensional case, discretization allows POMCP to outperform all other methods and DESPOT to perform well, but in subsequent problems where the observation space has more dimensions, discretization does not provide the same performance improvement (see Section 4.2.4).

4.3.3 Sub Hunt

In the Sub Hunt domain, the agent is a submarine attempting to track and destroy an enemy sub. The state and action spaces are discrete so that QMDP can be used

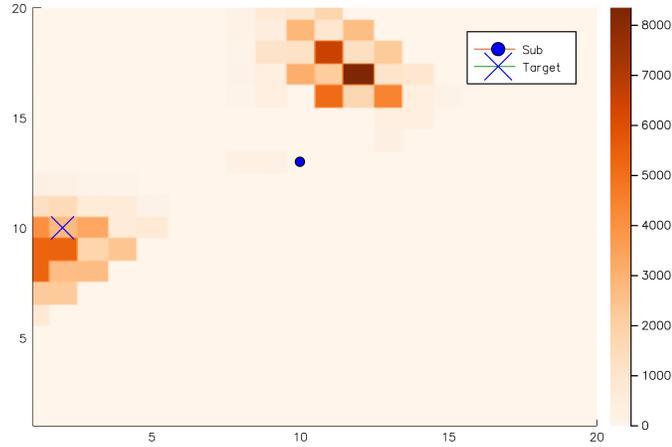


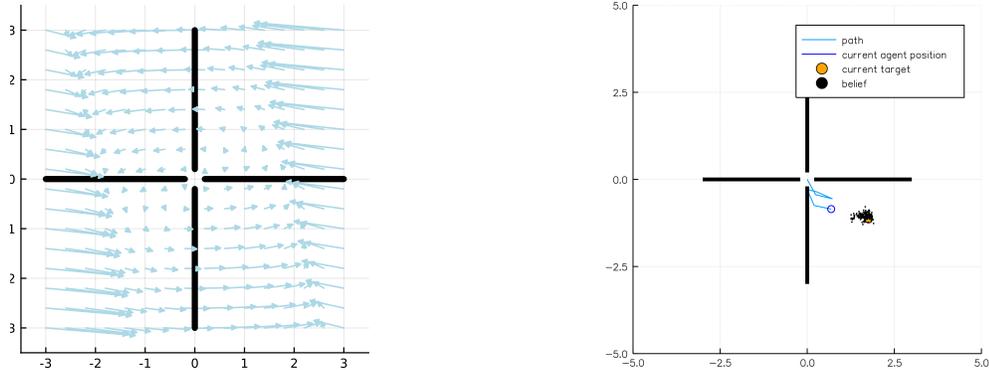
Figure 4.5: Sub Hunt benchmark. The color of each location square shows the number of belief particles containing the target at that location. In this case, the sub has incorrectly concluded that, with high probability, the target is moving towards the east and is chasing that belief concentration.

to solve the problem for comparison. The agent and the target each occupy a cell of a 20 by 20 grid. The target is either aware or unaware of the agent and seeks to reach a particular edge of the grid unknown to the agent ($\mathcal{S} = \{1, \dots, 20\}^4 \times \{\text{aware}, \text{unaware}\} \times \{N, S, E, W\}$). The target stochastically moves either two steps towards the goal or one step forward and one to the side. The agent has six actions, move three steps north, south, east, or west, engage the other submarine, or ping with active sonar. If the agent chooses to engage and the target is unaware and within a range of 2, a hit with reward 100 is scored; The problem ends when a hit is scored or the target reaches its goal edge.

An observation consists of 8 sonar returns ($\mathcal{O} = \mathbb{R}^8$) at equally-spaced angles that give a normally distributed estimate ($\sigma = 0.5$) of the range to the target if the target is within that beam and a measurement with higher variance if it is not. The range of the sensors depends on whether the agent decides to use active sonar. If the agent does not use active sonar it can only detect the other submarine within a radius of 3, but pinging with active sonar will detect at any range. However, active sonar alerts the target to the presence of the agent, and when the target is aware, the hit probability when engaging drops to 60%.

Table 4.1 shows the mean reward for 1000 simulations for each solver. The optimal strategy includes using the active sonar, but previous approaches have difficulty determining this because of the reduced engagement success rate. The PFT-DPW approach has the best score, followed closely by POMCPOW. All other solvers have similar performance to QMDP.

4.3.4 Van Der Pol Tag



(a) Van Der Pol Dynamics. The arrows show the target differential equation, and the thick black lines represent the barriers.

(b) Van Der Pol Tag. The agent attempts to tag the target, but cannot pass through barriers.

Figure 4.6: Van Der Pol tag problem

The final experimental problem is called Van Der Pol tag and has continuous state, action, and observation spaces. In this problem an agent moves through 2D space to try to tag a target ($\mathcal{S} = \mathbb{R}^4$) that has a random unknown initial position in $[-4, 4] \times [-4, 4]$. The agent always travels at the same speed, but chooses a direction of travel and whether to take an accurate observation ($\mathcal{A} = [0, 2\pi) \times \{0, 1\}$). The observation again consists of 8 beams ($\mathcal{O} = \mathbb{R}^8$) that give measurements to the target. Normally, these measurements are too noisy to be useful ($\sigma = 5$), but, if the agent chooses an accurate measurement with a cost of 5, the observation has low noise ($\sigma = 0.1$). The agent is blocked if it comes into contact with one of the barriers that stretch from 0.2 to 3.0 in each of the cardinal directions (see Fig. 4.6), while the target can move freely through. There is a cost of 1 for each step, and a reward of

100 for tagging the target (being within a distance of 0.1).

The target moves following a two dimensional form of the Van Der Pol oscillation defined by the differential equations

$$\dot{x} = \mu \left(x - \frac{x^3}{3} - y \right) \quad \text{and} \quad \dot{y} = \frac{1}{\mu} x,$$

where $\mu = 2$. Gaussian noise ($\sigma = 0.05$) is added to the position at the end of each step. Runge-Kutta fourth order integration is used to propagate the state.

This problem has several challenging features that might be faced in real-world applications. First, the state transitions are more computationally expensive because of the numerical integration. Second, the continuous state space and obstacles make it difficult to construct a good heuristic rollout policy, so random rollouts are used. Table 4.1 shows the mean reward for 1000 simulations of this problem for each solver. Since a POMCPOW iteration requires less computation than a PFT-DPW iteration, POMCPOW simulates more random rollouts and thus performs slightly better.

4.3.5 Multilane

The final problem for evaluation is the multiple-lane-change problem described in Section 3.2. Since costly information gathering is not an important part of this problem, POMCP-DPW and POMCPOW have similar performance. Because it uses a fixed number of scenarios and bounds to control exploration rather than progressive widening and a UCB heuristic, DESPOT is able to make better long term plans and is the most effective in this problem. It is also important to note that in this problem, the most realistic of the test problems, PFT-DPW performs significantly worse than POMCPOW.

4.3.6 Discretization granularity

Figure 4.7 shows the performance at different discretization granularities for the Light Dark and Sub Hunt problems.

Since the Light Dark domain has only a single observation dimension, it is easy

to discretize. In fact, POMCP with fine discretization outperforms POMCPOW. However, discretization is only effective at certain granularities, and this is highly dependent on the solver and possibly hyperparameters. In the Sub Hunt problem, with its high-dimensional observation, discretization is not effective at any granularity. In Van Der Pol tag, both the action and observation spaces must be discretized. Due to the high dimensionality of the observation space, similar to Sub Hunt, no discretization that resulted in good performance was found.

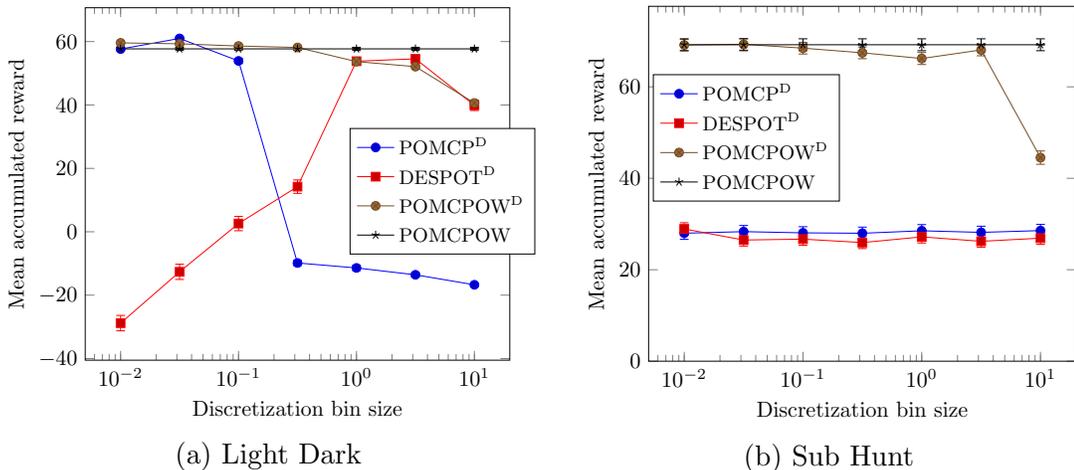


Figure 4.7: Discretization granularity studies

4.3.7 Hyperparameters

Hyperparameters for POMCPOW and PFT-DPW were chosen using the cross entropy method [66], but exact tuning was not a high priority and some parameters were re-used across solvers so the parameters may not be perfectly optimized. The values used in the experiments are shown in Table 4.2. There are not enough experiments to draw broad conclusions about the hyperparameters, but it appears that performance is most sensitive to the exploration constant, c .

The values for the observation widening parameters, k_o and α_o , were similar for all the problems in this work. A small α_o essentially limits the number of observations to a static number k_o , resulting in behavior reminiscent of sparse UCT [17], preventing unnecessary widening and allowing the tree to grow deep. This seems to work well in

Table 4.2: Hyperparameters used in experiments

POMCPOW	Laser Tag	Light Dark	Sub Hunt	VDP Tag	Multilane
c	26.0	90.0	17.0	110.0	8.0
k_a	–	–	–	30.0	–
α_a	–	–	–	1/30	–
k_o	4.0	5.0	6.0	5.0	4.5
α_o	1/35	1/15	1/100	1/100	1/10
PFT-DPW	Laser Tag	Light Dark	Sub Hunt	VDP Tag	Multilane
m	20	20	20	20	15
c	26.0	100.0	100.0	70.0	8.0
k_a	–	–	–	20.0	–
α_a	–	–	–	1/25	–
k_o	4.0	4.0	2.0	8.0	4.5
α_o	1/35	1/10	1/10	1/85	1/10

For problems with discrete actions, all actions are considered and k_a and α_a are not needed.

practice with the branching factor (k_o) set to values between 2 and 8, and suggests that it may be sufficient to limit the number of children to a fixed number rather than do progressive widening in a real implementation.

4.4 Discussion

The results in this chapter show that the previous leading online solvers are unable to adequately solve POMDPs with continuous observation spaces where exploration is important. In particular, Theorem 1 explains why these methods behave similarly to QMDP. POMCPOW and PFT-DPW overcome this limitation and outperform QMDP. Though POMCPOW only performs strictly best on one of the problems in Section 4.3, it performs very well on all of them, so it might be considered the best overall algorithm in these tests.

However there are still several shortcomings. First, there is no analytical proof given that POMCPOW or PFT-DPW converge toward the optimal solution. Second,

double progressive widening is made more difficult in practice because of its tuning parameters. Third, the results from the Multilane experiment show that DESPOT is able to find significantly better plans in more realistic problems. Thus, much work remains along this trajectory. DESPOT has several attractive attributes compared to the new algorithms presented here. In particular, its use of a fixed number of scenarios to control widening, and bounds to control exploration may make it more amenable to analysis and allow it to perform better in practice. Extending DESPOT to handle continuous observation spaces seems to be a promising future direction.

In order to make this type of research much easier, the Stanford Intelligent Systems Lab has created a software framework called POMDPs.jl which provides a standard interface for expressing problems and tools for rapidly prototyping and evaluating new algorithms using the Julia programming language. This framework is the subject of the next chapter.

Chapter 5

A Julia Framework for POMDPs

Since exact optimal solutions to POMDPs can rarely be attained, most research into solving realistic problems like those investigated in Chapters 2 and 3 involves empirical comparison between solution techniques. Sharing solver software between researchers can greatly improve the speed and quality of this work. Moreover, a consistent and concise framework for representing problems and demonstrating solution techniques is invaluable for POMDP education. This chapter describes the POMDPs.jl software package created by the Stanford Intelligent Systems Lab (SISL) to make state-of-the-art POMDP solution methods easily accessible to students, researchers, and engineers [32]. All of the research in Chapters 3 and 4 was conducted using this framework.

The POMDPs.jl package itself is implementation-free, and contains only the interface, however the JuliaPOMDP organization maintains several other packages that provide tools and concrete implementations. For example, the POMDPToolbox package contains a variety of useful tools for representing common distributions, beliefs, policies, etc. and the POMDPModels package contains some simple problem implementations.

5.1 Challenges for POMDP-solving software

A successful POMDP software framework must have, at a minimum, the following attributes: speed, flexibility, and ease of use. Achieving all of these attributes simultaneously is a major challenge.

5.1.1 Speed

Since POMDPs are difficult to solve [76], any computational slowdown such as unnecessary memory allocation or runtime type inference significantly reduces the maximum problem size that the framework can handle. For this reason, POMDP algorithms must be compiled to efficient processor instructions with low overhead.

5.1.2 Flexibility

The set of problems that can be represented as a POMDP is extremely large and there are many possible characteristics that such problems might have. A good POMDP software framework should try to accommodate as much of this set as possible. A few of the most important model characteristics to support are outlined below.

Partial and full observability

When studying a POMDP problem, it is almost always important to analyze the underlying fully-observable problem. Thus, a good POMDP framework should have first-class support for MDPs in addition to POMDPs.

Continuous and discrete problems

Some POMDPs have a finite number of states, actions, and observations, i.e. $|\mathcal{S}| < \infty$, $|\mathcal{A}| < \infty$, and $|\mathcal{O}| < \infty$. However, many real world problems, notably robotics problems, are naturally formulated in spaces with uncountably infinite cardinality, e.g. $\mathcal{S} = \mathbb{R}$, $\mathcal{A} = \mathbb{R}$, and $\mathcal{O} = \mathbb{R}$, multi-dimensional vector spaces, e.g. $\mathcal{S} = \mathbb{R}^6$, or hybrid continuous-discrete spaces. This means that the framework must not be

constrained to use integers for state representation, but should be capable of using a range of structures including floating point numbers and arrays.

Explicit vs generative model representation

Some POMDP and MDP solution techniques use the explicit probability distributions \mathcal{T} and \mathcal{Z} to solve problems. Thus, a successful framework must include a way to explicitly specify \mathcal{T} and \mathcal{Z} . On the other hand, explicitly specifying \mathcal{T} and \mathcal{Z} for many realistic problems is exceedingly difficult and tedious, and specifying a generative model is the only practical way to encode the problem. Thus, a successful framework must also include generative model support.

Online and offline solvers

While some POMDP solution techniques seek exact offline solutions to small problems, many larger problems can only be practically solved online. Thus a good POMDP framework must have first class support for solving offline and efficiently executing a policy online or executing a planner that does significant computation online.

Policy representation

The policies that different solution techniques yield can take a variety of forms. Exact solution techniques typically attempt to find alpha vectors that encode an optimal policy [47, 58], whereas others use finite state machines [8]. Newer methods may use neural networks [49] or other structures to store policies, so a successful framework must provide a flexible way to represent all of these structures.

5.1.3 Ease of Use

In addition to being flexible and fast, the framework must be easy to use. It is possible to make a framework that is performant and flexible but so complex that it will not be adopted or will cause much time to be wasted in understanding and implementation.

For the framework to be considered a success, it must be adopted by the community and serve as an enabler rather than a hindrance to research.

5.2 Previous frameworks

A number of frameworks exist for solving sequential decision problems. However, most frameworks are written from a reinforcement learning perspective and hence only support either fully observable problems or environments that can generate observations but do not provide access to the Markov state structure of the problem. Examples from this class are RLPy [35], and rllab [30]. The most closely related frameworks to POMDPs.jl are APPL [59], AI-Toolbox [9], BURLAP [25], and ZMDP [96] in that they explicitly represent the state and partial observability.

Of these POMDP frameworks, APPL contains the most advanced algorithms. It is written efficiently in C++ and is excellent from a speed perspective. However, it has several shortcomings in terms of flexibility and ease of use. First, though all of the solvers in APPL support the POMDPX file format for discrete, explicit problem definitions, flexibility is limited because there is not a unified interface for defining generative models or continuous explicit models. Second, prototyping problems and solvers in this framework is relatively time consuming and complex, making it difficult to use. Specifically, the POMDPX file format is based on XML and is difficult for humans to write directly, so, in most cases, custom scripts must be written to create the files, and solvers and generative models written in C++ have higher development time costs than those written in a higher level language. Third the main solvers implemented in APPL all have subtle interaction differences. In fact, the APPL website¹ reports “At the moment, the three [APPL solver] packages are independent. Maybe one day they will be merged in a single coherent framework.”

Though much research progress has been made with these frameworks, POMDPs.jl offers several significant improvements.

¹Permanent link: <https://perma.cc/5CKT-GC96>

5.3 Architecture

POMDPS.jl is designed to facilitate communication between different people performing three actions: defining problems, writing solver software, and running simulation experiments. The same person will often operate in two or even all three of these roles, but will nearly always use some code written by others.

The framework derives its solutions to the challenges outlined above primarily through the use of the Julia language itself [11]. Julia is just-in-time compiled using the state-of-the-art LLVM compiler framework [62], giving it speed comparable to traditional compiled languages such as C and C++. However, unlike other compiled languages, it has many features that make numerical software development easier and faster. For example, like Python, it is dynamically typed, has a powerful and flexible type system, and uses a modern, convenient syntax with features like list comprehensions. Moreover, like Matlab, it has built-in efficient multidimensional arrays and linear algebra. This makes meeting the flexibility and ease-of-use goals much easier.

The following section describes some of the concepts used in the framework before the interface itself is outlined.

5.3.1 Concepts

To fulfill each of the roles mentioned above, a programmer implements one or more classes that concretely represent concepts used to describe POMDPS and solvers. The problem writer creates a concrete subtype of the `POMDP` or `MDP` abstract types to represent a problem, the simulator writer creates a `Simulator` type to run simulations, and the solver writer creates a `Solver` subtype to run computations offline, and a `Policy` subtype to execute a policy online.

The most abstract concept in POMDPS.jl is the concept of belief. In the framework, the term “Belief” is used to mean any structure that encodes the information needed to execute the policy. For instance, if the policy is encoded as a set of alpha vectors, the belief takes its usual meaning as an explicit probability distribution over the states. For an online solver like POMCP, the belief must merely generate states

for the tree search. For a finite-state-machine policy like the one created by MCVI, the belief is simply the id of the current node. An `Updater` in POMDPs.jl is an object that defines how information from new observations is integrated into the belief. For example, if the belief is a probability distribution, the updater would apply Bayes rule or an approximation. Since the belief is often closely related to the policy, the solver writer will often implement the `Updater`, however generic updaters such as particle filters are also available. Figure 5.1 shows the three role concepts and some of the associated abstract types and interface functions.

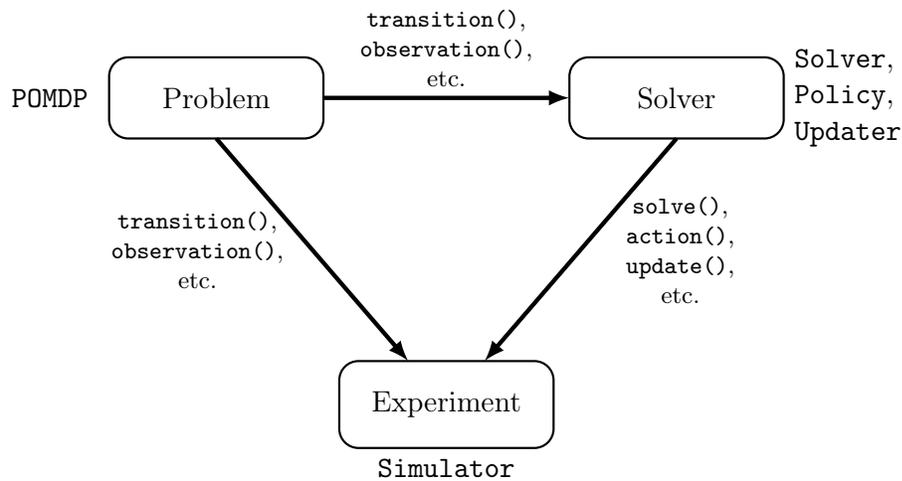


Figure 5.1: POMDPs.jl concepts. POMDPs.jl facilitates communication between people in three roles. The abstract types are shown beside each node and some of the interface functions are shown between the nodes. The arrows indicate which roles use code from which other roles.

5.3.2 Interfaces

The behavior of POMDPs.jl objects is defined by implementing methods of interface functions. Implementing interface functions serves as an alternative to writing configuration files in formats like POMDPX or specification scripts in purpose-built languages like RDDDL [87]. For example, the problem writer may implement a method of the `reward` function that returns the reward given a problem instance, state and action. Julia will call the correct `reward` method for the problem type based on its

multiple dispatch system [11]. Similarly, an `Updater` should have a corresponding `update` method that returns a new belief given an updater object, previous belief, action, and observation. The complete interface is not listed here, but is available in the online documentation.

In POMDPs.jl, states, actions, observations, beliefs, and distributions can be represented by any type as long as the appropriate interface methods are implemented. This provides the flexibility needed to represent continuous or discrete problems. Packages in the Julia ecosystem provide convenient and efficient types for common state representations, such as small fixed-size vectors.

One flexibility goal that has not been met by most previous frameworks is support of both generative and explicit problem definitions. For example, in APPL, all solvers can handle POMDPX problem specification for explicit definitions, but the MCVI and DESPOT solvers use different interfaces for generative problems, and there is no way to represent continuous problems explicitly. POMDPs.jl overcomes this challenge by exposing both an *explicit interface* and a *generative interface* that can be mixed. If the necessary parts of the explicit interface are implemented for a problem, POMDPs.jl will automatically provide the generative interface functions for the problem. The interface relationships are illustrated in Fig. 5.2. Implementations of all of the interface functions are not strictly required; if the minimum set of functions used by a solver or simulator are implemented for a problem, then that solver or simulator will work with that problem.

Because of the interface’s flexibility, giving users helpful error messages and feedback about which interface functions they should implement and even checking whether a sufficient portion of the interface has been implemented is a complex challenge. For example, suppose a user intends to implement a complex problem that cannot easily be expressed with an explicit definition and intends to use a solver that only requires functions from the generative interface. If POMDPs.jl advises this user to implement functions from the explicit interface, he or she will conclude that expressing the problem is difficult or impossible. This complexity drove the development of an interface and framework for dynamically specifying requirements and dependencies and generating helpful reports for users. In this requirements framework, built with

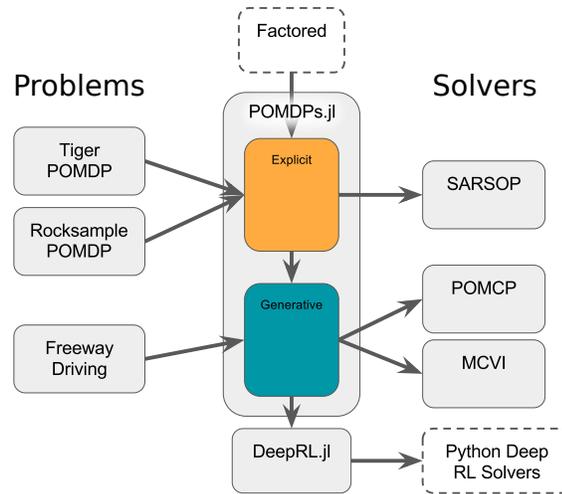


Figure 5.2: POMDPs.jl interfaces

Julia’s powerful metaprogramming tools, solver and simulator writers declare the requirements for their algorithm, which may be based on solver options, and problem writers can check which of the requirements have been satisfied by their problem implementation. Examples of the output of this system are shown in Fig. 5.3.

5.4 Examples

This section presents simple examples illustrating implementations of the concepts in Section 5.3.1, namely the problem, the solver, and the experiment.

5.4.1 Problem

The problem defines the MDP or POMDP model to be solved. For this example, consider the Tiger POMDP example of Kaelbling, Littman, and Cassandra [47]. An implementation of the type representing this POMDP is shown in Listing 1. The `TigerPOMDP` type is a subtype of the abstract `POMDP` type. The three parameters of `POMDP` denote the types used to represent the states, actions, and observations. These can be any types including built-in Julia types or special-purpose user defined types. Here `Symbols` (interned strings) are used for the sake of readability, but `Bools`, `Ints`,

```

julia> type NewMDP <: MDP{Int, Int} end

julia> @requirements_info ValueIterationSolver() NewMDP()

INFO: POMDPs.jl requirements for solve(::ValueIterationSolver, ::Union{POMDP,MDP}) and dependencies. ([✓]
= implemented correctly; [X] = missing)

For solve(::ValueIterationSolver, ::Union{POMDP,MDP}):
[✓] discount(::NewMDP)
[X] n_states(::NewMDP)
[X] n_actions(::NewMDP)
[X] transition(::NewMDP, ::Int64, ::Int64)
[X] reward(::NewMDP, ::Int64, ::Int64, ::Int64)
[✓] state_index(::NewMDP, ::Int64)
WARNING: Some requirements may not be shown because a MethodError was thrown.
For ordered_states(::Union{POMDP,MDP}) (in solve(::ValueIterationSolver, ::Union{POMDP,MDP})):
[X] states(::NewMDP)
WARNING: Some requirements may not be shown because a MethodError was thrown.
For ordered_actions(::Union{POMDP,MDP}) (in solve(::ValueIterationSolver, ::Union{POMDP,MDP})):
[✓] action_index(::NewMDP, ::Int64)
[X] actions(::NewMDP)
WARNING: Some requirements may not be shown because a MethodError was thrown.

Throwing the first exception (from processing solve(::ValueIterationSolver, ::Union{POMDP,MDP}) requireme
nts):

ERROR: MethodError: no method matching actions(::NewMDP)
Closest candidates are:
  actions{S}(::POMDPs.MDP{S,Int64}, ::S, ::Range{T}) at /home/zach/.julia/v0.5/POMDPToolbox/src/convenience/implem
entations.jl:11
  actions{S,A}(::Union{POMDPs.MDP{S,A},POMDPs.POMDP{S,A,0}}, ::S) at /home/zach/.julia/v0.5/POMDPs/src/sp
ace.jl:43
  actions{S,0}(::POMDPs.POMDP{S,Bool,0}) at /home/zach/.julia/v0.5/POMDPToolbox/src/convenience/implem
entations.jl:6
  ...
  in macro expansion at /home/zach/.julia/v0.5/DiscreteValueIteration/src/vanilla.jl:92 [inlined]
  in macro expansion at /home/zach/.julia/v0.5/POMDPs/src/requirements_internals.jl:51 [inlined]
  in get_requirements(::POMDPs.#solve, ::Tuple{DiscreteValueIteration.ValueIterationSolver,NewMDP}) at /ho
me/zach/.julia/v0.5/POMDPs/src/requirements_interface.jl:62
  in requirements_info(::DiscreteValueIteration.ValueIterationSolver, ::NewMDP) at /home/zach/.julia/v0.5/
POMDPs/src/requirements_interface.jl:140

```

(a) New problem with no methods

```

julia> using POMDPModels

julia> @requirements_info ValueIterationSolver() GridWorld()

INFO: POMDPs.jl requirements for solve(::ValueIterationSolver, ::Union{POMDP,MDP}) and dependencies. ([✓]
= implemented correctly; [X] = missing)

For solve(::ValueIterationSolver, ::Union{POMDP,MDP}):
[✓] discount(::GridWorld)
[✓] n_states(::GridWorld)
[✓] n_actions(::GridWorld)
[✓] transition(::GridWorld, ::GridWorldState, ::GridWorldAction)
[✓] reward(::GridWorld, ::GridWorldState, ::GridWorldAction, ::GridWorldState)
[✓] state_index(::GridWorld, ::GridWorldState)
[✓] iterator(::GridWorldActionSpace)
[✓] iterator(::GridWorldStateSpace)
[✓] iterator(::GridWorldDistribution)
[✓] pdf(::GridWorldDistribution, ::GridWorldState)
For ordered_states(::Union{POMDP,MDP}) (in solve(::ValueIterationSolver, ::Union{POMDP,MDP})):
[✓] states(::GridWorld)
For ordered_actions(::Union{POMDP,MDP}) (in solve(::ValueIterationSolver, ::Union{POMDP,MDP})):
[✓] action_index(::GridWorld, ::GridWorldAction)
[✓] actions(::GridWorld)

```

(b) Fully implemented problem

Figure 5.3: POMDPs.jl requirements example output for the value iteration solver

or a custom enumeration could also easily be used.

The problem behavior is defined by implementing new methods of the functions in the POMDPs.jl interface. For example, the `observation` function in Listing 1 returns the observation distribution for the Tiger POMDP given an action and state. The `SparseCat` object returned by this method is provided by the POMDPToolbox package, represents a (potentially sparse) categorical distribution over a list of objects. Other code can randomly sample from the `SparseCat` distribution, query its probability distribution, and efficiently iterate over the members with nonzero probability. Since the state in the Tiger POMDP never changes, the transition dynamics can be expressed concisely as shown in Listing 1.

Listing 1 Partial implementation of the Tiger POMDP

```

struct TigerPOMDP <: POMDP{Symbol, Symbol, Symbol}
    p_correct::Float64 # probability of hearing the tiger correctly
    discount::Float64 # discount factor
end

function POMDPs.observation(pomdp::TigerPOMDP, a::Symbol, s::Symbol)
    pc = pomdp.p_correctly
    if a == :listen
        if s == :tiger_left
            p = pomdp.p_correct
        else
            p = 1.0 - pomdp.p_correct
        end
    else
        p = 0.5
    end
    return SparseCat([:tiger_left, :tiger_right], [p, 1.0-p])
end

function POMDPs.transition(pomdp::TigerPOMDP, s::Symbol, a::Symbol)
    if a == :listen
        return SparseCat([s], [1.0])
    else # a looking action - the problem finishes
        return SparseCat([:done], [1.0])
    end
end

```

The entire problem definition contains several more methods, including `reward`, `actions`, `discount`, `isterminal`, etc.

5.4.2 Solver

A solver implementation requires between one and three type definitions: a `Solver` that contains the parameters that define solver behavior, often a `Policy` that defines a mapping from beliefs to actions, and sometimes an `Updater` if the belief is closely linked to how the policy functions. Some example code for the QMDP [65] solution method is shown below.

QMDP works by running value iteration on the fully observable MDP at the core. In the `solve` function, partially shown in Listing 2, the `POMDPS.jl` functions which will have methods for the particular POMDP being solved are called. Once value iteration has completed, an `AlphaVectorPolicy` from `POMDPToolbox.jl` is returned. The ellipses indicate code that has been left out for the sake of brevity.

Listing 2 Partial implementation of the QMDP solution technique

```
function POMDPS.solve(solver::QMDPSolver, pomdp::POMDP)
    ...
    for i in 1:solver.max_iterations
        ...
        for (istate,s) in enumerate(states)
            if isterminal(mdp, s)
                util[istate] = 0.0
                ...
            else
                ...
                for a in iterator(actions(pomdp, s))
                    dist = transition(mdp, s, a)
                    u = 0.0
                    for (sp, p) in weighted_iterator(dist)
                        r = reward(pomdp, s, a, sp)
                        isp = state_index(pomdp, sp)
                        u += p * (r + discount(pomdp) * util[isp])
                    end
                end
                ...
            end
        end
    end
    return AlphaVectorPolicy(qvalues)
end
```

5.4.3 Simulator

POMDP solution algorithms are typically evaluated by running Monte Carlo simulations. A `Simulator` calls `POMDPS.jl` functions defined for the solver and the problem

to run a simulation and return the desired statistics from that simulation. For example, the experimenter might create a simulator type, `Sim`, and a corresponding `simulate` function (an important part of the main simulation loop is shown in Listing 3).

Listing 3 Partial implementation of a simulation routine

```
function simulate(sim::Sim,
                 pomdp::POMDP,
                 policy::Policy,
                 filter=updater(policy))
    ...
    for t in 1:sim.max_steps
        a = action(policy, b)
        sp = rand(sim.rng, transition(pomdp, s, a))
        r_total += discount(pomdp)^t*reward(pomdp, s, a, sp)
        o = rand(sim.rng, observation(pomdp, s, a, sp))
        b = update(filter, b, a, o)
        ...
    end
end
```

With the examples above and the appropriate additional methods defined, a simulation can then be run as shown in Listing 4.

Listing 4 Program to solve the Tiger POMDP and simulate it

```
pomdp = TigerPOMDP()           # initialize the tiger problem
solver = QMDPSolver()          # initialize QMDP solver
policy = solve(solver, pomdp)  # compute a policy
r = simulate(Sim(), pomdp, policy)
```

5.5 Discussion

By using the leverage of the Julia programming language, POMDPS.jl provides an unprecedented level of flexibility for expressing and solving POMDPs. It is being used by other researchers to investigate topics as diverse as GPS jammer detection by drones and cancer screening recommendations, along with numerous student projects for a course taught at Stanford.

However, there is still much work to be done. First, for the simplest examples, POMDPs.jl is a rather clumsy interface in that it requires methods to be implemented to define behavior. Some of the best interfaces for expressing mathematical optimization problems, such as JuMP [31] and CVX [40], have concise, declarative interfaces that can express problems in a few lines of code and be easily interpreted by looking at one page for a short time. The POMDP domain is much more diverse and complex than the optimization domains addressed by these packages, but a concise declarative interface that can represent a subset of POMDPs would be very helpful in the education mission of POMDPs.jl.

Second, many of the solvers are far from achieving the full performance potential of the algorithms that they implement. Since most modern hardware performance increases are due to parallelism rather than single-process speed increases, one of the most important factors in speeding up solver code is parallel implementation. Julia has basic built-in tools for parallel processing, but there are still several pitfalls for implementing parallel code. For example, since modern processors use caching extensively to improve memory access, efficient parallel code should only access memory that is likely to be in the local cache and not being used by another processor [27]. An investment of effort into parallelizing solver code would make the POMDPs.jl ecosystem much more useful.

Chapter 6

Summary and Future Work

In order to realize benefits of autonomous transportation, the control systems of autonomous vehicles must be safe. However, if this safety brings too great a sacrifice in efficiency, the vehicles will not be adopted as widely or be as useful. This thesis investigated approaches for improving safety while sacrificing minimal efficiency, specifically by proposing better methods for handling the uncertainty inherent in the environment autonomous vehicles operate in. In addition, the thesis proposed improved algorithms for solving the optimization problems that arise from vehicle control problems. Finally, it described the software framework created to make education and research collaboration in this area easier. The remainder of this chapter reviews the specific contributions of this thesis and outlines potential future research directions.

6.1 Contributions and Summary

The first contribution (Chapter 2) is an investigation into the **combination of trusted resolution logic and optimization** for UAV collision avoidance. Simple TRL has the advantage that guarantees about its operation are relatively easy to certify. However, a comparison between simple TRL alone and MDP optimization shows that there is a price in terms of both safety and efficiency accrued for using the

TRL. To reduce this price, two methods for combining the approaches are investigated. The first is to use an MDP policy to dynamically adjust the parameters of the TRL that govern its conservativeness. The second is to use the TRL as a constraint on the actions that the MDP policy can choose. Simulation experiments show that the second approach is superior.

The second contribution (Chapter 3) is a **study of uncertainty modeling in decision making for autonomous highway lane changing**. This chapter considered approximate POMDP solutions where the internal states of other drivers are explicitly modeled as a hidden part of the Markov state. The POMDP solutions are compared with approximate MDP solutions, both assuming full knowledge of the internal state and conservatively modeling state uncertainty as outcome uncertainty. **The advantage of POMDP solutions over MDP solutions is clear** in all of the tests, but the relative effectiveness of different POMDP approximations depends greatly on the correlation of the different internal state parameters. Simulation tests show that when the parameters are highly correlated, since it is easier to estimate the hidden states, even greatly simplified POMDP approximations can achieve performance near the upper bound established by an omnipotent planner. On the other hand, when there is little or no correlation, POMDP solution methods that include more of the uncertainty in the solution process perform significantly better. Experiments also characterize the robustness of the algorithms with respect to the parameter distribution.

The third contribution (Chapter 4) is a pair of **improved algorithms for solving POMDPs online**. Previous leading online algorithms have focused on solving discrete problems, and are unable to solve problems with continuous action and observation spaces. Hence, new algorithms for the continuous domains encountered in the real world are needed. The first part of this contribution (Section 4.2.1) is proving that **a naive application of double progressive widening to the POMCP algorithm will result in suboptimal behavior**. In particular, in a continuous observation space, it will converge to the QMDP solution because each belief node will degenerate to a single state particle (Theorem 1). In fact, numerical experiments confirm that all leading online solvers exhibit this behavior.

In response to this suboptimality, two new algorithms are proposed (Sections 4.2.2 and 4.2.3). The first, PFT-DPW, solves an approximation of the belief MDP, while the second, **POMCPOW**, is an extension of POMCP with double progressive widening and weighted particle filtering. Numerical experiments (Section 4.3) show that these algorithms are able to break the QMDP barrier and hence exhibit performance that is qualitatively superior to previous algorithms. POMCPOW and PFT-DPW perform similarly on most problems, but on the most realistic problem, a version of the autonomous driving problem from Chapter 3, POMCPOW outperforms PFT-DPW.

The final contribution (Chapter 5) is a software framework, **POMDPs.jl**, which provides an interface for expressing MDPs and POMDPs and tools for developing and testing solvers. This framework is unprecedented in its flexibility. It can represent discrete and continuous MDPs defined by both explicit probability distributions and simulators that provide generative models. The Julia language also gives solvers written to use the framework speed comparable to previous C++ implementations.

6.2 Future Work

There are many promising directions for future work to proceed from the results presented in this thesis. The discussion portion of each chapter gives short-term specific extensions to the individual research efforts described therein. This section gives a broader overview.

The first line of future work involves **safety constraints**. Interaction of safety constraints with sequential optimization is explored in Chapter 2 and such constraints are used in Chapter 3. However, the actual constraints used there are very simple. In real problems, these constraints will need to be more complex, consisting of, for example, linear temporal logic rules [83] or results from reachability analysis [21]. Combining approximate POMDP solutions with these advanced safety constraints to field a reliable system still requires much research.

The second line is focused on **modeling**. Chapters 2 and 3 evaluate their findings based on simplistic models of other vehicles. To solidify the conclusions of this thesis, better data-driven models of human drivers and pilots must be created. Indeed,

significant work has already been done in this area [88, 12]. However, these models are limited in scope, primarily because of the limited data available. In order to be able to test in more diverse scenarios, more data will need to be collected from the real world, and models that are more data-efficient developed. Moreover, current planning algorithms will not work efficiently with certain models. A particular example is that recurrent neural networks with latent states that represent beliefs over internal states [88] will not work with the POMDP algorithms used in this thesis. Thus, new algorithms must be developed to work with the new models.

The third direction is continuing the development of better **POMDP algorithms**. While the algorithms proposed in Chapter 4 demonstrate better performance than previous algorithms, the shortcomings of this family of algorithms are also apparent. In particular, algorithms based on MCTS-DPW have many tuning parameters that must be chosen via unreliable hand tuning or time consuming optimization. In addition, the proof of consistency for MCTS-DPW [4] is rather tedious, which is one of the reasons that an analytical proof for the optimality of POMCPOW was not attempted for this thesis. Moreover, the trees produced by these algorithms tend to be shallow, reducing the effective planning horizon. As seen in the experimental results for the multilane model in Table 4.1, DESPOT is able to perform better on more realistic problems because it builds deeper trees and explores based on bounds rather than upper confidence estimates. DESPOT’s fixed number of scenarios may also make analysis easier. Given these advantages, the particle filter weighting approaches used in PFT-DPW and POMCPOW should be applied to more advanced algorithms like DESPOT to enable them to explore in continuous observation spaces. Furthermore, all of the algorithms discussed in this thesis use a single execution thread. In order to take advantage of modern computing hardware, online parallel algorithms for solving POMDPs must be investigated.

Finally, the problem of controlling autonomous vehicles in continuous, partially observable domains has by no means been solved. One reason is the weakness of decision-making **algorithms for continuous spaces** that can handle uncertainty and irregularity. This thesis focused on continuous state and observation spaces, but the action space is arguably the most difficult context for continuity. Some research

has begun to address this [90, 109], but these algorithms use only derivative-free optimization techniques that may not scale to high dimensional problems. In many continuous problems, if history h_1 consists of actions and observations near the actions and observations of history h_2 , the value at h_1 will be very close to the value at h_2 . This information is not exploited in any of the online solvers discussed in this thesis. Although some work has been done to use this information in MCTS [111], new algorithms that use ideas from optimal control, model predictive control, and motion planning may be much better suited for continuous problems.

Appendix A

Proof of Theorem 1

A version of Monte Carlo tree search with double progressive widening has been proven to converge to the optimal value function on fully observable MDPs by Auger, Couetoux, and Teytaud [4]. We use this proof to show that POMCP-DPW converges to the QMDP solution.

First we establish some preliminary definitions taken directly from Auger, Couetoux, and Teytaud [4].

Definition 2 (Regularity Hypothesis). *The Regularity hypothesis is the assumption that for any $\Delta > 0$, there is a non zero probability to sample an action that is optimal with precision Δ . More precisely, there is a $\theta > 0$ and a $p > 1$ (which remain the same during the whole simulation) such that for all $\Delta > 0$,*

$$Q(ha) \geq Q^*(h) - \Delta \text{ with probability at least } \min(1, \theta\Delta^p). \quad (\text{A.1})$$

Definition 3 (Exponentially sure in n). *We say that some property depending on an integer n is exponentially sure in n if there exists positive constants C , h , and η such that the probability that the property holds is at least*

$$1 - C \exp(-hn^\eta).$$

In order for the proof from Auger, Couetoux, and Teytaud [4] to apply, the following four minor modifications to the POMCP-DPW algorithm must be made:

1. Instead of the usual logarithmic exploration, use *polynomial exploration*, that is, select actions based on the criterion

$$Q(ha) + \sqrt{\frac{N(h)^{e_d}}{N(ha)}}, \quad (\text{A.2})$$

as opposed to the traditional criterion

$$Q(ha) + c\sqrt{\frac{\log N(h)}{N(ha)}}, \quad (\text{A.3})$$

and create a new node for progressive widening when $\lfloor N^\alpha \rfloor > \lfloor (N-1)^\alpha \rfloor$ rather than when the number of children exceeds kN^α .

2. Instead of performing rollout simulations, keep creating new single-child nodes until the maximum depth is reached.
3. In line 15, instead of selecting an observation randomly, select the observation that has been visited least proportionally to how many times it has been generated.
4. Use the depth-dependent coefficient values in Table 1 from Auger, Couetoux, and Teytaud [4] instead of choosing static values.

This version of the algorithm will be referred to as “modified POMCP-DPW”. The algorithm with these changes is listed in Algorithm 6.

We now define the “QMDP value” that POMCP-DPW converges to (this is repeated from the main text of the paper) and prove a preliminary lemma.

Definition 1 (QMDP value). *Let $Q_{MDP}(s, a)$ be the optimal state-action value function assuming full observability starting by taking action a in state s . The QMDP value at belief b , $Q_{MDP}(b, a)$, is the expected value of $Q_{MDP}(s, a)$ when s is distributed according to b .*

Lemma 1. *If POMCP-DPW or modified POMCP-DPW is applied to a POMDP with a continuous observation space and observation probability density functions that are finite everywhere, then each history node in the tree will have only one corresponding state, that is $|B(h)| = 1, M(h) = 1 \forall h$.*

Proof. Since the observation probability density function is finite, each call to the generative model will produce a unique observation with probability 1. Because of this, lines 18 and 19 of Algorithm 6 will only be executed once for each observation. \square

We are now ready to restate the theorem from the text.

Theorem 1 (Modified POMCP-DPW convergence to QMDP). *If a bounded-horizon POMDP meets the following conditions: 1) the state and observation spaces are continuous with a finite observation probability density function, and 2) the regularity hypothesis is met, then modified POMCP-DPW will produce a value function estimate, \hat{Q} , that converges to the QMDP value for the problem. Specifically, there exists a constant $C > 0$, such that after n iterations,*

$$\left| \hat{Q}(b, a) - Q_{MDP}(b, a) \right| \leq \frac{C}{n^{1/(10d_{\max}-7)}}$$

exponentially surely in n , for every action a .

The bound on the value estimate error, $\frac{C}{n^{1/(10d_{\max}-7)}}$, is based on the specific coefficients chosen by Auger, Couetoux, and Teytaud [4] and listed in Table 1 of their paper. Alternative bounds may be possible with different coefficient choices. The proof of the theorem is given below.

Proof. We prove that modified POMCP-DPW functions exactly as the Polynomial UCT (PUCT) algorithm defined by Auger, Couetoux, and Teytaud [4] applied to an augmented fully observable MDP, and hence converges to the QMDP value. We will show this by proposing incremental changes to Algorithm 6 that do not change its function that will result in an algorithm identical to PUCT.

Before listing the changes, we define the “augmented fully observable MDP” as follows: For a POMDP $\mathcal{P} = (\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \mathcal{O}, \mathcal{Z}, \gamma)$, and belief b , the *augmented fully observable MDP*, \mathcal{M} , is the MDP defined by $(\mathcal{S}_A, \mathcal{A}, \mathcal{T}_A, \mathcal{R}, \gamma)$, where

$$\mathcal{S}_A = \mathcal{S} \cup \{b\} \quad (\text{A.4})$$

and, for all $x, x' \in \mathcal{S}_A$,

$$\mathcal{T}_A(x'|x, a) = \begin{cases} \mathcal{T}(x'|x, a) & \text{if } x \in \mathcal{S} \\ \int_{\mathcal{S}} b(s) \mathcal{T}(x'|s, a) ds & \text{if } x = b \end{cases} \quad (\text{A.5})$$

This is simply the fully observable MDP augmented with a special state representing the current belief. It is clear that the value function for this problem, $Q_{\mathcal{M}}(b, a)$, is the same as the QMDP value for the POMDP, $Q_{\text{MDP}}(b, a)$. Thus, by showing that modified POMCP-DPW behaves exactly as PUCT applied to \mathcal{M} , we show that it estimates the QMDP values.

Consider the following modifications to Algorithm 6 that do not change its behavior when the observation space is continuous:

1. Eliminate the state count M . *Justification:* By Lemma 1, its value will be 1 for every node.
2. Remove B and replace with a mapping H from each node to a state of \mathcal{M} ; define $H(b) = b$. *Justification:* By Lemma 1, B always contains only a single state, so H contains the same information.
3. Generate states and rewards with $G_{\mathcal{M}}$, the generative model of \mathcal{M} , instead of G . *Justification:* Since the state transition model for the fully observable MDP is the same as the POMDP, these are equivalent for all $s \in \mathcal{S}$.
4. Remove the s argument of SIMULATE. *Justification:* The sampling in line 3 is done implicitly in $G_{\mathcal{M}}$ if $h = b$, and s is redundant in other cases because h can be mapped to s through H .

The result of these changes is shown in Algorithm 7. It is straightforward to verify that this algorithm is equivalent to PUCT applied to \mathcal{M} . Each observation-terminated history, h , corresponds to a PUCT “decision node”, z , and each action-terminated history, ha , corresponds to a PUCT “chance node”, w . In other words, the observations have no meaning in the tree other than making up the histories, which are effectively just keys to or aliases for the state nodes.

Since PUCT is guaranteed by Theorem 1 of Auger, Couetoux, and Teytaud [4] to converge to the optimal value function of \mathcal{M} exponentially surely, POMCP-DPW is guaranteed to converge to the QMDP value exponentially surely, and the theorem is proven. □

Remark 1. *One may object that multiple histories may map to the same state through H , and thus the history nodes in a modified POMCP-DPW tree are not equivalent to state nodes in the PUCT tree. In fact, the PUCT algorithm does not check to see if a state has previously been generated by the model, so it may also contain multiple decision nodes z that correspond to the same state. Though this is not explicitly stated by the authors, it is clear from the algorithm description, and the proof still holds.*

Algorithm 6 Modified POMCP-DPW

```

1: procedure PLAN( $b$ )
2:   for  $i \in 1 : n$ 
3:      $s \leftarrow$  sample from  $b$ 
4:     SIMULATE( $s, b, d_{\max}$ )
5:   return  $\operatorname{argmax}_a Q(ba)$ 

6: procedure ACTIONPROGWIDEN( $h$ )
7:   if  $\lfloor N(h)^{\alpha_{a,d}} \rfloor > \lfloor (N(h) - 1)^{\alpha_{a,d}} \rfloor$ 
8:      $a \leftarrow$  NEXTACTION( $h$ )
9:      $C(h) \leftarrow C(h) \cup \{a\}$ 
10:  return  $\operatorname{argmax}_{a \in C(h)} Q(ha) + \sqrt{\frac{N(h)^{\varepsilon_d}}{N(ha)}}$ 

11: procedure SIMULATE( $s, h, d$ )
12:  if  $d = 0$ 
13:    return 0
14:   $a \leftarrow$  ACTIONPROGWIDEN( $h$ )
15:  if  $\lfloor N(ha)^{\alpha_{o,d}} \rfloor > \lfloor (N(ha) - 1)^{\alpha_{o,d}} \rfloor$ 
16:     $s', o, r \leftarrow G(s, a)$ 
17:     $C(ha) \leftarrow C(ha) \cup \{o\}$ 
18:     $M(hao) \leftarrow M(hao) + 1$ 
19:    append  $s'$  to  $B(hao)$ 
20:  else
21:     $o \leftarrow \operatorname{argmin}_{o \in C(ha)} N(hao)/M(hao)$ 
22:     $s' \leftarrow$  select  $s' \in B(hao)$  w.p.  $\frac{1}{|B(hao)|}$ 
23:     $r \leftarrow \mathcal{R}(s, a, s')$ 
24:   $total \leftarrow r + \gamma \text{SIMULATE}(s', hao, d - 1)$ 
25:   $N(h) \leftarrow N(h) + 1$ 
26:   $N(ha) \leftarrow N(ha) + 1$ 
27:   $Q(ha) \leftarrow Q(ha) + \frac{total - Q(ha)}{N(ha)}$ 
28:  return  $total$ 

```

Algorithm 7 Modified POMCP-DPW on a continuous observation space

```

1: procedure PLAN( $b$ )
2:   for  $i \in 1 : n$ 
3:     SIMULATE( $(b), d_{\max}$ )
4:   return  $\operatorname{argmax}_a Q(ha)$ 

5: procedure ACTIONPROGWIDEN( $h$ )
6:   if  $\lfloor N(h)^{\alpha_{a,d}} \rfloor > \lfloor (N(h) - 1)^{\alpha_{a,d}} \rfloor$ 
7:      $a \leftarrow \text{NEXTACTION}(h)$ 
8:      $C(h) \leftarrow C(h) \cup \{a\}$ 
9:   return  $\operatorname{argmax}_{a \in C(h)} Q(ha) + \sqrt{\frac{N(h)^{e_d}}{N(ha)}}$ 

10: procedure SIMULATE( $h, d$ )
11:   if  $d = 0$ 
12:     return 0
13:    $a \leftarrow \text{ACTIONPROGWIDEN}(h, d)$ 
14:   if  $\lfloor N(ha)^{\alpha_{o,d}} \rfloor > \lfloor (N(ha) - 1)^{\alpha_{o,d}} \rfloor$ 
15:      $\cdot, o, \cdot \leftarrow G(H(h), a)$ 
16:      $H(hao), r \leftarrow G_{\mathcal{M}}(H(h), a)$ 
17:      $C(ha) \leftarrow C(ha) \cup \{o\}$ 
18:   else
19:      $o \leftarrow \operatorname{argmin}_{o \in C(ha)} N(hao)$ 
20:      $r \leftarrow \mathcal{R}(H(h), a, H(hao))$ 
21:    $total \leftarrow r + \gamma \text{SIMULATE}(hao, d - 1)$ 
22:    $N(h) \leftarrow N(h) + 1$ 
23:    $N(ha) \leftarrow N(ha) + 1$ 
24:    $Q(ha) \leftarrow Q(ha) + \frac{total - Q(ha)}{N(ha)}$ 
25:   return  $total$ 

```

Bibliography

- [1] A.-A. Agha-Mohammadi, S. Chakravorty, and N. M. Amato, “FIRM: Feedback controller-based information-state roadmap - a framework for motion planning under uncertainty,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2011 (cit. on p. 16).
- [2] E. Altman, *Constrained Markov decision processes*. CRC Press, 1999 (cit. on pp. 6, 10).
- [3] M. Araya, O. Buffet, V. Thomas, and F. Charpillet, “A POMDP extension with belief-dependent rewards,” in *Advances in Neural Information Processing Systems (NIPS)*, 2010 (cit. on p. 81).
- [4] D. Auger, A. Couetoux, and O. Teytaud, “Continuous upper confidence trees with polynomial exploration–consistency,” in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, 2013, pp. 194–209 (cit. on pp. 78, 79, 109, 111–113, 115).
- [5] R. Bach, C. Farrell, and H. Erzberger, “An Algorithm for Level-Aircraft Conflict Resolution,” NASA, Tech. Rep. CR-2009-214573, 2009 (cit. on pp. 20, 21).
- [6] H. Bai, S. Cai, N. Ye, D. Hsu, and W. S. Lee, “Intention-aware online POMDP planning for autonomous driving in a crowd,” in *IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 454–460 (cit. on p. 49).
- [7] H. Bai, D. Hsu, M. J. Kochenderfer, and W. S. Lee, “Unmanned aircraft collision avoidance using continuous-state POMDPs,” in *Robotics: Science and Systems*, 2012 (cit. on pp. 20, 21, 31).

- [8] H. Bai, D. Hsu, and W. S. Lee, “Integrated perception and planning in the continuous space: A POMDP approach,” *International Journal of Robotics Research*, vol. 33, no. 9, pp. 1288–1302, 2014 (cit. on pp. 15, 95).
- [9] E. Bargiacchi, *AI-Toolbox*, <https://github.com/Svalorzen/AI-Toolbox> (cit. on p. 96).
- [10] D. Bertsekas, *Dynamic Programming and Optimal Control*. Athena Scientific, 2005 (cit. on pp. 9, 10, 12, 32–34, 36).
- [11] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, “Julia: A fresh approach to numerical computing,” *SIAM Review*, vol. 59, no. 1, pp. 65–98, 2017 (cit. on pp. 97, 99).
- [12] R. Bhattacharyya, D. P. Phillips, B. Wulfe, J. Morton, A. Kuefler, and M. J. Kochenderfer, “Multi-agent imitation learning for driving simulation,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018 (cit. on p. 109).
- [13] J. A. Boyan and A. W. Moore, “Generalization in reinforcement learning: Safely approximating the value function,” in *Advances in Neural Information Processing Systems*, 1995, pp. 369–376 (cit. on p. 37).
- [14] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004 (cit. on p. 63).
- [15] S. Brechtel, T. Gindele, and R. Dillmann, “Solving continuous POMDPs: Value iteration with incremental learning of an efficient space representation,” in *International Conference on Machine Learning (ICML)*, 2013 (cit. on p. 49).
- [16] S. Brechtel, T. Gindele, and R. Dillmann, “Solving continuous POMDPs: Value iteration with incremental learning of an efficient space representation,” in *International Conference on Machine Learning (ICML)*, 2013, pp. 370–378 (cit. on p. 15).
- [17] C. B. Browne, E. Powley, D. Whitehouse, *et al.*, “A survey of Monte Carlo tree search methods,” *IEEE Transactions on Computational Intelligence and AI in games*, vol. 4, no. 1, pp. 1–43, 2012 (cit. on pp. 12–14, 49, 90).

- [18] A. Bry and N. Roy, “Rapidly-exploring random belief trees for motion planning under uncertainty,” in *IEEE International Conference on Robotics and Automation (ICRA)*, 2011, pp. 723–730 (cit. on p. 16).
- [19] M. Buehler, K. Iagnemma, and S. Singh, Eds., *The DARPA urban challenge: Autonomous vehicles in city traffic*, ser. Tracts in Advanced Robotics. Springer, 2010, vol. 56, p. 628 (cit. on p. 3).
- [20] B. Casey, “Amoral machines, or: How roboticists can learn to stop worrying and love the law,” *Northwestern University Law Review*, vol. 111, p. 1347, 2016 (cit. on p. 4).
- [21] M. Chen and C. J. Tomlin, “Exact and efficient Hamilton-Jacobi reachability for decoupled systems,” in *IEEE Conference on Decision and Control (CDC)*, 2015, pp. 1297–1303 (cit. on p. 108).
- [22] Y. Chow, A. Tamar, S. Mannor, and M. Pavone, “Risk-sensitive and robust decision-making: A CVaR optimization approach,” in *Advances in Neural Information Processing Systems*, 2015, pp. 1522–1530 (cit. on p. 9).
- [23] A. Couëtoux, J.-B. Hoock, N. Sokolovska, O. Teytaud, and N. Bonnard, “Continuous upper confidence trees,” in *Learning and Intelligent Optimization*, Rome, Italy, 2011 (cit. on pp. 14, 22, 74).
- [24] S. Davies, “Multidimensional triangulation and interpolation for reinforcement learning,” in *Advances in Neural Information Processing Systems*, 1996, pp. 1005–1011 (cit. on p. 36).
- [25] C. Diuk, A. Cohen, and M. L. Littman, “An object-oriented representation for efficient reinforcement learning,” in *International Conference on Machine Learning*, ACM, 2008, pp. 240–247 (cit. on p. 96).
- [26] D. Dolgov, *Google self-driving car project monthly report*, Google, Inc., Sep. 2016 (cit. on pp. 3, 50).
- [27] U. Drepper, *What every programmer should know about memory*, <https://www.akkadia.org/drepper/cpumemory.pdf>, Red Hat, Inc, 2007 (cit. on p. 105).

- [28] L. Dressel and M. Kochenderfer, “Efficient decision-theoretic target localization,” in *International Conference on Automated Planning and Scheduling (ICAPS)*, 2017 (cit. on p. 81).
- [29] K. Driggs-Campbell and R. Bajcsy, “Identifying modes of intent from driver behaviors in dynamic environments,” in *IEEE International Conference on Intelligent Transportation Systems (ITSC)*, 2015 (cit. on p. 49).
- [30] Y. Duan, X. Chen, R. Houthoofd, J. Schulman, and P. Abbeel, “Benchmarking deep reinforcement learning for continuous control,” *ArXiv:1604.06778*, 2016 (cit. on p. 96).
- [31] I. Dunning, J. Huchette, and M. Lubin, “JuMP: A modeling language for mathematical optimization,” *SIAM Review*, vol. 59, no. 2, pp. 295–320, 2017 (cit. on p. 105).
- [32] M. Egorov, Z. N. Sunberg, E. Balaban, T. A. Wheeler, J. K. Gupta, and M. J. Kochenderfer, “POMDPs.jl: A framework for sequential decision making under uncertainty,” *Journal of Machine Learning Research*, vol. 18, no. 26, pp. 1–5, 2017 (cit. on pp. 19, 93).
- [33] J. Foster and F. B. Richards, “The Gibbs phenomenon for piecewise-linear approximation,” *The American Mathematical Monthly*, vol. 98, no. 1, pp. 47–49, 1991 (cit. on p. 33).
- [34] C. E. Garcia, D. M. Prett, and M. Morari, “Model predictive control: Theory and practice – a survey,” *Automatica*, vol. 25, no. 3, pp. 335–348, 1989 (cit. on p. 58).
- [35] A. Geramifard, C. Dann, R. H. Klein, W. Dabney, and J. P. How, “RLPy: A value-function-based reinforcement learning framework for education and research,” *Journal of Machine Learning Research*, vol. 16, pp. 1573–1578, 2015 (cit. on p. 96).
- [36] T. Gindele, S. Brechtel, and R. Dillmann, “Learning driver behavior models from traffic observations for decision making and planning,” *IEEE Intelligent Transportation Systems Magazine*, vol. 7, no. 1, pp. 69–79, 2015 (cit. on p. 49).

- [37] *Global status report on road safety*, World Health Organization, 2015. [Online]. Available: http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/ (cit. on pp. 1, 2).
- [38] A. Goldhoorn, A. Garrell, R. Alquézar, and A. Sanfeliu, “Continuous real time POMCP to find-and-follow people by a humanoid service robot,” in *IEEE-RAS International Conference on Humanoid Robots*, 2014 (cit. on p. 73).
- [39] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. Cambridge: MIT Press, 2016 (cit. on pp. 33, 47).
- [40] M. Grant and S. Boyd, *CVX: Matlab software for disciplined convex programming, version 2.1*, <http://cvxr.com/cvx>, Mar. 2014 (cit. on p. 105).
- [41] G. Hagen, R. Butler, and J. Maddalon, “Stratway: A modular approach to strategic conflict resolution,” in *AIAA Aviation Technology, Integration, and Operations (ATIO) Conference*, 2011 (cit. on pp. 20, 21).
- [42] H. Herencia-Zapana, J.-B. Jeannin, and C. Muñoz, “Formal verification of safety buffers for sate-based conflict detection and resolution,” in *International Congress of the Aeronautical Sciences*, 2010 (cit. on pp. 20, 21).
- [43] J. Hoey and P. Poupart, “Solving POMDPs with continuous or large discrete observation spaces,” in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2005, pp. 1332–1338 (cit. on p. 16).
- [44] J. E. Holland, M. J. Kochenderfer, and W. A. Olson, “Optimizing the next generation collision avoidance system for safe, suitable, and acceptable operational performance,” *Air Traffic Control Quarterly*, vol. 21, no. 3, pp. 275–297, 2013 (cit. on pp. 20, 21).
- [45] V. Indelman, L. Carlone, and F. Dellaert, “Planning in the continuous domain: A generalized belief space approach for autonomous navigation in unknown environments,” *International Journal of Robotics Research*, vol. 34, no. 7, pp. 849–882, 2015 (cit. on p. 16).
- [46] R. Johnsonbaugh and W. E. Pfaffenberger, *Foundations of mathematical analysis*. Dover, 2010 (cit. on p. 12).

- [47] L. P. Kaelbling, M. L. Littman, and A. Cassandra, “Planning and acting in partially observable stochastic domains,” *Artificial Intelligence*, vol. 101, pp. 99–134, 1998 (cit. on pp. 10, 15, 95, 100).
- [48] L. P. Kaelbling and T. Lozano-Pérez, “Finding aircraft collision-avoidance strategies using policy search methods,” MIT-CSAIL, Tech. Rep. TR-2009-043, 2009 (cit. on pp. 20, 21).
- [49] P. Karkus, D. Hsu, and W. S. Lee, “QMDP-net: Deep learning for planning under partial observability,” in *Advances in Neural Information Processing Systems*, 2017, pp. 4697–4707 (cit. on pp. 17, 95).
- [50] A. Kesting, M. Treiber, and D. Helbing, “General lane-changing model MOBIL for car-following models,” *Transportation Research Record*, vol. 1999, pp. 86–94, 2007 (cit. on pp. 52, 71).
- [51] ———, “Agents for traffic simulation,” in *Multi Agent Systems: Simulation and Applications*, A. M. Uhrmacher and D. Weyns, Eds., CRC Press, 2009, ch. 11, pp. 325–356 (cit. on pp. 55, 60).
- [52] M. J. Kochenderfer, *Decision making under uncertainty: Theory and application*. MIT Press, 2015 (cit. on pp. 9, 10, 12, 46, 79).
- [53] M. J. Kochenderfer and J. P. Chryssanthacopoulos, “Robust airborne collision avoidance through dynamic programming,” MIT Lincoln Laboratory, Tech. Rep. ATC-371, 2011 (cit. on pp. 20, 21, 24, 31).
- [54] M. J. Kochenderfer, J. P. Chryssanthacopoulos, L. P. Kaelbling, and T. Lozano-Pérez, “Model-based optimization of airborne collision avoidance logic,” MIT Lincoln Laboratory, Tech. Rep. ATC-360, 2010 (cit. on pp. 20, 21).
- [55] M. J. Kochenderfer, J. P. Chryssanthacopoulos, and P. P. Radecki, “Robustness of optimized collision avoidance logic to modeling errors,” in *Digital Avionics Systems Conference*, 2010 (cit. on pp. 24, 42).

- [56] M. J. Kochenderfer, M. W. M. Edwards, L. P. Espindle, J. K. Kuchar, and J. D. Griffith, “Airspace encounter models for estimating collision risk,” *AIAA Journal of Guidance, Control, and Dynamics*, vol. 33, no. 2, pp. 487–499, 2010 (cit. on p. 24).
- [57] J. K. Kuchar and L. C. Yang, “A review of conflict detection and resolution modeling methods,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 1, no. 4, pp. 179–189, 2000 (cit. on p. 20).
- [58] H. Kurniawati, D. Hsu, and W. S. Lee, “SARSOP: Efficient point-based POMDP planning by approximating optimally reachable belief spaces.,” in *Robotics: Science and Systems*, Zurich, Switzerland., 2008 (cit. on pp. 15, 95).
- [59] H. Kurniawati, D. Hsu, W. S. Lee, and H. Bai, *Approximate POMDP Planning Toolkit*, <http://bigbird.comp.nus.edu.sg/pmwiki/farm/appl/> (cit. on p. 96).
- [60] H. Kurniawati and V. Yadav, “An online POMDP solver for uncertainty planning in dynamic environment,” in *Robotics Research*, Springer, 2016, pp. 611–629 (cit. on p. 17).
- [61] C.-P. Lam, A. Y. Yang, K. Driggs-Campbell, R. Bajcsy, and S. S. Sastry, “Improving human-in-the-loop decision making in multi-mode driver assistance systems using hidden mode stochastic hybrid systems,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015 (cit. on p. 49).
- [62] C. Lattner and V. Adve, “LLVM: A compilation framework for lifelong program analysis & transformation,” in *Proceedings of the International Symposium on Code Generation and Optimization*, IEEE Computer Society, 2004, p. 75 (cit. on p. 97).
- [63] J. R. Lepird, M. P. Owen, and M. J. Kochenderfer, “Bayesian preference elicitation for multiobjective engineering design optimization,” *Journal of Aerospace Information Systems*, vol. 12, no. 10, pp. 634–645, 2015 (cit. on p. 6).

- [64] S. Levine and V. Koltun, “Continuous inverse optimal control with locally optimal examples,” in *International Conference on Machine Learning (ICML)*, 2012, pp. 41–48 (cit. on p. 6).
- [65] M. L. Littman, A. R. Cassandra, and L. P. Kaelbling, “Learning policies for partially observable environments: Scaling up,” in *International Conference on Machine Learning (ICML)*, 1995 (cit. on pp. 17, 79, 103).
- [66] S. Mannor, R. Rubinfeld, and Y. Gat, “The cross entropy method for fast policy search,” in *International Conference on Machine Learning (ICML)*, 2003, pp. 512–519 (cit. on p. 90).
- [67] B. McKenzie, “Who drives to work? Commuting by automobile in the United States: 2013,” United States Census Bureau, Tech. Rep. ACS-32, Aug. 2015 (cit. on p. 2).
- [68] *Mean travel time to work of workers 16 years and over who did not work at home*, <https://factfinder.census.gov/>, United States Census Bureau, 2016 (cit. on p. 2).
- [69] N. A. Melchior and R. Simmons, “Particle RRT for path planning with uncertainty,” in *IEEE International Conference on Robotics and Automation (ICRA)*, 2007 (cit. on p. 16).
- [70] V. Mnih, K. Kavukcuoglu, D. Silver, *et al.*, “Human-level control through deep reinforcement learning,” *Nature*, vol. 518, 2015 (cit. on p. 33).
- [71] P. Morere, R. Marchant, and F. Ramos, “Bayesian optimisation for solving continuous state-action-observation POMDPs,” in *Advances in Neural Information Processing Systems (NIPS)*, 2016 (cit. on p. 15).
- [72] A. Narkawicz, C. Muñoz, and G. Dowek, “Provably correct conflict prevention bands algorithms,” *Science of Computer Programming*, vol. 77, no. 10–11, pp. 1039–1057, 2012 (cit. on pp. 20, 21).
- [73] K. Naughton, “Humans are slamming into driverless cars and exposing a key flaw,” 2015. [Online]. Available: <http://bloom.bg/1Qw8fjB> (cit. on p. 50).
- [74] R. B. Nelsen, *An introduction to copulas*. Springer, 2007 (cit. on p. 60).

- [75] H. Y. Ong and M. J. Kochenderfer, “Short-term conflict resolution for unmanned aircraft traffic management,” in *Digital Avionics Systems Conference*, 2015 (cit. on p. 20).
- [76] C. H. Papadimitriou and J. N. Tsitsiklis, “The complexity of Markov decision processes,” *Mathematics of Operations Research*, vol. 12, no. 3, pp. 441–450, 1987 (cit. on pp. 11, 94).
- [77] A. Pas, “Simulation based planning for partially observable markov decision processes with continuous observation spaces,” Master’s thesis, Maastricht University, 2012 (cit. on p. 75).
- [78] R. Platt Jr., R. Tedrake, L. Kaelbling, and T. Lozano-Perez, “Belief space planning assuming maximum likelihood observations,” in *Robotics: Science and Systems*, 2010 (cit. on p. 15).
- [79] S. Prentice and N. Roy, “The belief roadmap: Efficient planning in belief space by factoring the covariance,” *International Journal of Robotics Research*, vol. 28, no. 11-12, pp. 1448–1465, 2009 (cit. on p. 16).
- [80] E. J. Rodríguez-Seda, “Decentralized Trajectory Tracking with Collision Avoidance Control for Teams of Unmanned Vehicles with Constant Speed,” in *American Control Conference*, 2014 (cit. on pp. 20, 24).
- [81] S. Ross, J. Pineau, S. Paquet, and B. Chaib-Draa, “Online planning algorithms for POMDPs,” *Journal of Artificial Intelligence Research*, vol. 32, pp. 663–704, 2008 (cit. on p. 16).
- [82] D. Sadigh, K. Driggs-Campbell, A. Puggelli, *et al.*, “Data-driven probabilistic modeling and verification of human driver behavior,” in *AAAI Spring Symposium on Formal Verification and Modeling in Human-Machine Systems*, 2014 (cit. on p. 49).
- [83] D. Sadigh and A. Kapoor, “Safe control under uncertainty with probabilistic signal temporal logic,” in *Robotics: Science and Systems*, 2016 (cit. on p. 108).

- [84] D. Sadigh, S. S. Sastry, S. A. Seshia, and A. Dragan, “Information gathering actions over human internal state,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016 (cit. on p. 49).
- [85] —, “Planning for autonomous cars that leverage effects on human actions,” in *Robotics: Science and Systems*, 2016 (cit. on p. 6).
- [86] G. Sandoval, “Alphabet says 2 of its ambitious side projects are ready for the big leagues,” *Business Insider*, Jul. 2018. [Online]. Available: <http://www.businessinsider.com/alphabet-projects-wing-loon-graduate-into-standalone-businesses-2018-7> (cit. on p. 3).
- [87] S. Sanner, “Relational dynamic influence diagram language (RDDL): Language description,” 2010, [Online]. Available: http://users.cecs.anu.edu.au/~ssanner/IPPC_2011/RDDL.pdf (cit. on p. 98).
- [88] E. Schmerling, K. Leung, W. Vollprecht, and M. Pavone, “Multimodal probabilistic model-based planning for human-robot interaction,” in *IEEE International Conference on Robotics and Automation (ICRA)*, Brisbane, Australia, May 2018 (cit. on pp. 71, 109).
- [89] B. Schoettle and M. Sivak, “A preliminary analysis of real-world crashes involving self-driving vehicles,” University of Michigan Transportation Research Institute, Tech. Rep. UMTRI-2015-34, 2015 (cit. on p. 49).
- [90] K. M. Seiler, H. Kurniawati, and S. P. N. Singh, “An online and approximate solver for POMDPs with continuous action space,” in *IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 2290–2297 (cit. on pp. 74, 110).
- [91] H. Shaban, “Amazon is issued patent for delivery drones that can react to screaming voices, flailing arms,” *The Washington Post*, Mar. 2018. [Online]. Available: <https://www.washingtonpost.com/news/the-switch/wp/2018/03/22/amazon-issued-patent-for-delivery-drones-that-can-react-to-screaming-flailing-arms> (cit. on p. 3).

- [92] D. Silver, A. Huang, C. J. Maddison, *et al.*, “Mastering the game of Go with deep neural networks and tree search,” *Nature*, vol. 529, no. 7587, pp. 484–489, 2016 (cit. on p. 16).
- [93] D. Silver and J. Veness, “Monte-Carlo planning in large POMDPs,” in *Advances in Neural Information Processing Systems (NIPS)*, 2010 (cit. on p. 83).
- [94] S. Singh, “Critical reasons for crashes investigated in the national motor vehicle crash causation survey,” National Highway Traffic Safety Administration, Tech. Rep. DOT HS 812 115, Feb. 2015. [Online]. Available: <https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812115> (cit. on p. 2).
- [95] *Small unmanned aircraft systems*, 14 C.F.R. §107, 2018 (cit. on p. 24).
- [96] T. Smith, *ZMDP software for POMDP planning*, <https://github.com/trey0/zmdp> (cit. on p. 96).
- [97] A. Somani, N. Ye, D. Hsu, and W. S. Lee, “DESPOT: Online POMDP planning with regularization,” in *Advances in Neural Information Processing Systems (NIPS)*, 2013, pp. 1772–1780 (cit. on pp. 17, 84).
- [98] Z. N. Sunberg, C. J. Ho, and J. Kochenderfer Mykel, “The value of inferring the internal state of traffic participants for autonomous freeway driving,” in *American Control Conference (ACC)*, 2017 (cit. on p. 19).
- [99] Z. N. Sunberg and M. J. Kochenderfer, “Online algorithms for POMDPs with continuous state, action, and observation spaces,” in *International Conference on Automated Planning and Scheduling (ICAPS)*, Delft, 2018 (cit. on p. 19).
- [100] Z. Sunberg, M. Kochenderfer, and M. Pavone, “Optimized and trusted collision avoidance for unmanned aerial vehicles using approximate dynamic programming,” in *Proc. IEEE Conf. on Robotics and Automation*, 2016 (cit. on p. 19).
- [101] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT Press, 1998 (cit. on p. 8).

- [102] S. Temizer, M. J. Kochenderfer, L. P. Kaelbling, T. Lozano-Perez, and J. K. Kuchar, “Collision avoidance for unmanned aircraft using markov decision processes,” in *AIAA Guidance, Navigation, and Control Conference*, 2010 (cit. on pp. 20, 21).
- [103] S. Thrun, “Monte Carlo POMDPs,” in *Advances in Neural Information Processing Systems (NIPS)*, vol. 12, 1999, pp. 1064–1070 (cit. on p. 15).
- [104] S. Thrun, W. Burgard, and D. Fox, *Probabilistic robotics*. MIT Press, 2005 (cit. on p. 14).
- [105] C. Tomlin, G. J. Pappas, and S. S. Sastry, “Conflict resolution for air traffic management: A study in multiagent hybrid systems,” *IEEE Transactions on Automatic Control*, vol. 43, no. 4, pp. 509–21, 1998 (cit. on p. 24).
- [106] M. Treiber, A. Hennecke, and D. Helbing, “Congested traffic states in empirical observations and microscopic simulations,” *Physical Review E*, vol. 62, no. 2, pp. 1805–1824, 2000 (cit. on pp. 52, 71).
- [107] C. Urmson, J. Anhalt, D. Bagnell, *et al.*, *Tartan racing: A multi-modal approach to the DARPA urban challenge*, <http://repository.cmu.edu/cgi/viewcontent.cgi?article=1967&context=robotics>, Carnegie Mellon University, 2007 (cit. on p. 4).
- [108] J. Van Den Berg, S. Patil, and R. Alterovitz, “Motion planning under uncertainty using iterative local optimization in belief space,” *International Journal of Robotics Research*, vol. 31, no. 11, pp. 1263–1278, 2012 (cit. on p. 16).
- [109] E. Wang, H. Kurniawati, and D. P. Kroese, “An online planner for POMDPs with large discrete action space: A quantile-based approach,” in *International Conference on Automated Planning and Scheduling (ICAPS)*, 2018 (cit. on pp. 74, 110).
- [110] T. A. Wheeler, P. Robbel, and M. J. Kochenderfer, “A probabilistic framework for microscopic traffic propagation,” in *IEEE International Conference on Intelligent Transportation Systems (ITSC)*, Las Palmas de Gran Canaria, Spain, 2015 (cit. on p. 49).

- [111] C. Xiao, J. Mei, and M. Müller, “Memory-augmented Monte Carlo tree search,” in *AAAI Conference on Artificial Intelligence (AAAI)*, 2018 (cit. on p. 110).
- [112] R. Zhang and M. Pavone, “Control of robotic mobility-on-demand systems: A queueing-theoretical perspective,” *International Journal of Robotics Research*, 2015 (cit. on p. 2).